Model Compression

Danna Gurari
University of Colorado Boulder
Spring 2024

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/
Review

• Last lecture:
  • Motivation
  • Overview of self-supervised learning
  • Generative-based methods
  • Generative adversarial networks
  • Context-based methods

• Assignments (Canvas):
  • Project outline due Wednesday

• Questions?
Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
Larger models perform best (with lots of training data):

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022
Modern Neural Networks Are a Mismatch for Many Real-World Applications


https://en.wikipedia.org/wiki/Wearable_technology

https://www.buzzfeednews.com/article/katienotopoulos/facebook-is-making-camera-glasses-ha-ha-oh-no
Modern Neural Networks Are a Mismatch for Many Real-World Applications

- **Large inference time** (i.e., incompatible for real-time applications)

- **Large memory footprint** (e.g., incompatible with limited memory on edge devices)

- **Large computational cost** (e.g., incompatible with limited battery on edge devices)

- Potential for **large environmental costs**
Idea: develop compact models so deep learning models can be used more efficiently and for more applications
Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
Popular Approach: Knowledge Distillation

A student learns from a knowledgeable teacher

Image source: https://www.waterford.org/education/teacher-student-relationships/
Key Question: What is Knowledge?

Parameters W!
Knowledge Is: Input to Output Mapping

Mapping: Input to Output!
Knowledge Is: Input to Output Mapping

Target mapping: ground truth (1-hot vector)
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers further insights into similarities and differences of categories
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that 2 has similar characteristics to 7 and 1
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that bear has similar characteristics to dog and cat
Knowledge Is: Input to Output Mapping

Target mapping: probability distribution from a model offers further insights into similarities and differences of categories
- Attempts to identify ground truth category
- Also, shares that bear has similar characteristics to dog and cat

Idea: teach about ground truth and its relationships to other categories

Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher

Loss computed to bring student’s distribution closer to that of the teacher
Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher

Also transfers knowledge about the relative similarity of other categories to the predicted category
Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

Recall **Softmax**: converts vector of scores into a probability distribution that sums to 1

![Diagram](https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjlxODk)

Want to enhance knowledge of this relationship
Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

Recall **Softmax**: converts vector of *scores* into a probability distribution that sums to 1

Get rid of negative values while preserving original order of scores

\[
\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}
\]

- \(i = 1, \ldots, K\)
- Number of classes
- Divide each node’s score by sum of all entries to make them sum to 1 (normalization)

Useful tutorial: https://towardsdatascience.com/exploring-the-softmax-function-578c8b0fb15
Knowledge Distillation: Rebalance ("Soften")
Probability Distribution Across Categories

Generalized Softmax: converts vector of scores into a probability distribution that sums to 1 with temperature

$$\sigma(z)_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

What is the typical value of T used for softmax?
Idea: set the temperature to a value greater than 1
Knowledge Distillation: Rebalance ("Soften")
Probability Distribution Across Categories

**Generalized Softmax**: converts vector of scores into a probability distribution that sums to 1 with temperature

\[
\sigma(z)_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}
\]

Larger T values mean more information is available about which categories the teacher found similar to the predicted category

[https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk](https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjkxODk)
Knowledge Distillation: Rebalance ("Soften") Probability Distribution Across Categories

**Generalized Softmax**: converts vector of scores into a probability distribution that sums to 1 with temperature; e.g.,

\[
\begin{bmatrix}
0.997 & 0.000 & 0.002 & 0.001 & 0.000 \\
0.935 & 0.0001 & 0.046 & 0.017 & 0.0001 \\
0.637 & 0.021 & 0.191 & 0.128 & 0.021 \\
\end{bmatrix}
\]

- **T=1**
- **T=2**
- **T=5**

Larger T values mean more information is available about which categories the teacher found similar to the predicted category

Knowledge Distillation: Rebalance (“Soften”) Probability Distribution Across Categories

**Generalized Softmax**: converts vector of scores into a probability distribution that sums to 1 with temperature; e.g., T=5

Knowledge of this relationship enhanced from 0.01

https://wandb.ai/authors/knowledge-distillation/reports/Distilling-Knowledge-in-Neural-Networks--VmlldzoyMjhxODk
Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher

Teacher (pre-trained model) → Student (to be trained) → “Distillation” loss computed to bring student’s distribution closer to that of the teacher, using the generalized softmax equation
Knowledge Distillation: Teach Student the “Dark Knowledge” of Teacher

Total loss computed during training is a weighted sum of the conventional cross entropy loss and the “distillation loss”.

Teacher (pretrained model)

Student (to be trained)

https://blog.csdn.net/qq_22749699/article/details/79460817
Knowledge Distillation: At Test Time

Student model
Arguably, Any Neural Network Student Could Learn from Any Neural Network Teacher

https://blog.csdn.netqq_22749699/article/details/79460817
Arguably, Any Neural Network Student Could Learn from Any Neural Network Teacher

Knowledge distillation is a type of transfer learning

https://blog.csdn.net/qq_22749699/article/details/79460817
Knowledge Distillation Enhancement: Hints

Encourage student (FitNet) to mimic the teacher’s feature responses; e.g., output of guided layer should match the output of hint layer.

Knowledge Distillation Enhancement: Hints

Encourage student (FitNet) to mimic the teacher’s feature responses; e.g., output of guided layer should match the output of hint layer

Training conducted to learn the intermediate feature

\[ w_{\text{Guided}} = \arg \min_{w_{\text{Guided}}} L_{HT}(w_{\text{Guided}}, w_{i}) \]

Layer added to match size of the hint’s output layer

---

(a) Teacher and Student Networks  
(b) Hints Training

Knowledge Distillation Enhancement: Hints

Encourage student (FitNet) to mimic the teacher’s feature responses; e.g., output of guided layer should match the output of hint layer.

After learning the intermediate features, the whole student network is trained.

(a) Teacher and Student Networks
(b) Hints Training
(c) Knowledge Distillation

Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
Recall Task: Predict Category from 1000 Options

- Evaluation metric: % correct (top-1 and top-5 predictions)
- Dataset: ~1.5 million images
- Source: images scraped from search engines, such as Flickr, and labeled by crowdworkers

Experiment: Do Bigger, More Accurate Models Make Better Teachers?

Tested Teachers: ResNet18/34/50/152

Trained Student: ResNet18

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019

Figure source: https://blog.csdn.net/qq_22749699/article/details/79460817
Experiment: Do Bigger, More Accurate Models Make Better Teachers?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Teacher Error (%)</th>
<th>Student Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>30.24</td>
<td>30.57</td>
</tr>
<tr>
<td>ResNet34</td>
<td>26.70</td>
<td>30.79</td>
</tr>
<tr>
<td>ResNet50</td>
<td>23.85</td>
<td>30.95</td>
</tr>
</tbody>
</table>

What is the student’s performance trend from larger, more accurate teachers?
Experiment: Do Bigger, More Accurate Models Make Better Teachers?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Teacher Error (%)</th>
<th>Student Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>30.24</td>
</tr>
<tr>
<td>ResNet18</td>
<td>30.24</td>
<td>30.57</td>
</tr>
<tr>
<td>ResNet34</td>
<td>26.70</td>
<td>30.79</td>
</tr>
<tr>
<td>ResNet50</td>
<td>23.85</td>
<td>30.95</td>
</tr>
</tbody>
</table>

Student performance not only drops for larger teachers but the models distilled from teachers perform worse than training the student from scratch!
Experiment: Why Might Student Performance Drop as Teacher Size Grows?

1. More accurate models are more confident and so need higher temperatures to learn the “dark knowledge” of category relationships

2. Student mimics teacher but the loss function is mismatched from the evaluation metric

3. Student fails to accurately mimic teacher

Experimental analysis suggests this is the reason.
Hypothesis: student is underfitting because of lower capacity and so “minimizing one loss (KD loss) at the expense of the other (cross entropy loss)”
Experiment: Why Might Students Fail to Mimic Teachers?

How to overcome this issue?

- Early stopping with KD loss (ESKD) to leverage its benefit at the start of training

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019
Experiments: How Does ESKD Compare To Training A Student from Scratch?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Top-1 Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>30.57</td>
</tr>
<tr>
<td>ResNet18 (ES KD)</td>
<td>29.01</td>
</tr>
<tr>
<td>ResNet34</td>
<td>30.79</td>
</tr>
<tr>
<td>ResNet34 (ES KD)</td>
<td>29.16</td>
</tr>
<tr>
<td>ResNet50</td>
<td>30.95</td>
</tr>
<tr>
<td>ResNet50 (ES KD)</td>
<td>29.35</td>
</tr>
</tbody>
</table>

Training a model with early stopping knowledge distillation loss leads to better results than training from scratch!
Experiments: Are Results from ESKD Better When Using Bigger, More Accurate Models As Teachers?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Top-1 Error (%, Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>30.57</td>
</tr>
<tr>
<td>ResNet18 (ES KD)</td>
<td>29.01</td>
</tr>
<tr>
<td>ResNet34</td>
<td>30.79</td>
</tr>
<tr>
<td>ResNet34 (ES KD)</td>
<td>29.16</td>
</tr>
<tr>
<td>ResNet50</td>
<td>30.95</td>
</tr>
<tr>
<td>ResNet50 (ES KD)</td>
<td>29.35</td>
</tr>
</tbody>
</table>

No; the student may still be struggling with underfitting due to an insufficient representational capacity

Cho and Hariharan. On the Efficacy of Knowledge Distillation. ICCV 2019
Experiments: To Address The Capacity Problem
Why Not Instead Distill to Intermediate Sizes?

Performs almost identically to a model that is distilled directly from a large to small size; does not address the core problem:

The student must be in the solution space of the teacher
Today’s Topics

- Motivation
- Key idea: knowledge distillation (KD)
- Pioneering KD model for image classification
- Pioneering KD model for object detection
- State-of-the-art for KD
Recall Popular Detection Model: Faster R-CNN

Approach for Creating Compact Student Model

Faster R-CNN

Teacher
- Hint
- L2 Loss
- Adaptation

Student
- Guided

Detection
- Classification
- Soft Label
- Regression
- Bounded Regression Loss
- Weighted Cross Entropy Loss

Predicts a category and parameters to adjust the size and location of the input BB

Ground Truth Label
- SoftMax & SmoothL1 Loss
- Ground Truth

= backpropagation pathways

A loss is computed to encourage the student’s intermediate features to match those of the teacher.

Classification distillation loss:
- Computed if the student's distance to the GT exceeds the teacher's distance.

Regression distillation loss:
- Computed if the student’s distance to the GT exceeds the teacher’s distance.

Conventional loss:
- Computed for classification and regression errors compared to the GT.

## Experiments

### mAP scores for 5 datasets

<table>
<thead>
<tr>
<th>Student</th>
<th>Model Info</th>
<th>Teacher</th>
<th>PASCAL</th>
<th>COCO@.5</th>
<th>COCO@[.5,.95]</th>
<th>KITTI</th>
<th>ILSVRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucker</td>
<td>11M / 47ms</td>
<td>-</td>
<td>54.7</td>
<td>25.4</td>
<td>11.8</td>
<td>49.3</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>- AlexNet</td>
<td>-</td>
<td>57.6 (+2.9)</td>
<td>26.5 (+1.2)</td>
<td>12.3 (+0.5)</td>
<td>51.4 (+2.1)</td>
<td>23.6 (+1.3)</td>
</tr>
<tr>
<td></td>
<td>- VGGM</td>
<td>-</td>
<td>58.2 (+3.5)</td>
<td>26.4 (+1.1)</td>
<td>12.2 (+0.4)</td>
<td>51.4 (+2.1)</td>
<td>23.9 (+1.6)</td>
</tr>
<tr>
<td></td>
<td>- VGG16</td>
<td>-</td>
<td>59.4 (+4.7)</td>
<td>28.3 (+2.9)</td>
<td>12.6 (+0.8)</td>
<td>53.7 (+4.4)</td>
<td>24.4 (+2.1)</td>
</tr>
<tr>
<td>AlexNet</td>
<td>62M / 74ms</td>
<td>- VGGM</td>
<td>57.2</td>
<td>32.5</td>
<td>15.8</td>
<td>55.1</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>- VGG16</td>
<td>-</td>
<td>59.2 (+2.0)</td>
<td>33.4 (+0.9)</td>
<td>16.0 (+0.2)</td>
<td>56.3 (+1.2)</td>
<td>28.7 (+1.4)</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>60.1 (+2.9)</td>
<td>35.8 (+3.3)</td>
<td>16.9 (+1.1)</td>
<td>58.3 (+3.2)</td>
<td>30.1 (+2.8)</td>
</tr>
<tr>
<td>VGGM</td>
<td>80M / 86ms</td>
<td>- VGG16</td>
<td>59.8</td>
<td>33.6</td>
<td>16.1</td>
<td>56.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>63.7 (+3.9)</td>
<td>37.2 (+3.6)</td>
<td>17.3 (+1.2)</td>
<td>58.6 (+2.3)</td>
<td>34.0 (+2.9)</td>
</tr>
<tr>
<td>VGG16</td>
<td>138M / 283ms</td>
<td>-</td>
<td>70.4</td>
<td>45.1</td>
<td>24.2</td>
<td>59.2</td>
<td>35.6</td>
</tr>
</tbody>
</table>

### What trends do you observe from these results?

# Experiments

For all student-teacher pairs, knowledge distillation yields more compact, faster, and more accurate detections.


<table>
<thead>
<tr>
<th>Student Model Info</th>
<th>Teacher Model Info</th>
<th>PASCAL</th>
<th>COCO@0.5</th>
<th>COCO@[0.5:0.95]</th>
<th>KITTI</th>
<th>ILSVRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucker 11M / 47ms</td>
<td>-</td>
<td>54.7</td>
<td>25.4</td>
<td>11.8</td>
<td>49.3</td>
<td>20.6</td>
</tr>
<tr>
<td>AlexNet</td>
<td>57.6 (+2.9)</td>
<td>26.5 (+1.2)</td>
<td>12.3 (+0.5)</td>
<td>51.4 (+2.1)</td>
<td>23.6 (+1.3)</td>
<td></td>
</tr>
<tr>
<td>VGGGM</td>
<td>58.2 (+3.5)</td>
<td>26.4 (+1.1)</td>
<td>12.2 (+0.4)</td>
<td>51.4 (+2.1)</td>
<td>23.9 (+1.6)</td>
<td></td>
</tr>
<tr>
<td>VGG16</td>
<td>59.4 (+4.7)</td>
<td>28.3 (+2.9)</td>
<td>12.6 (+0.8)</td>
<td>53.7 (+4.4)</td>
<td>24.4 (+2.1)</td>
<td></td>
</tr>
<tr>
<td>AlexNet 62M / 74ms</td>
<td>-</td>
<td>57.2</td>
<td>32.5</td>
<td>15.8</td>
<td>55.1</td>
<td>27.3</td>
</tr>
<tr>
<td>VGGGM</td>
<td>59.2 (+2.0)</td>
<td>33.4 (+0.9)</td>
<td>16.0 (+0.2)</td>
<td>56.3 (+1.2)</td>
<td>28.7 (+1.4)</td>
<td></td>
</tr>
<tr>
<td>VGG16</td>
<td>60.1 (+2.9)</td>
<td>35.8 (+3.3)</td>
<td>16.9 (+1.1)</td>
<td>58.3 (+3.2)</td>
<td>30.1 (+2.8)</td>
<td></td>
</tr>
<tr>
<td>VGGM 80M / 86ms</td>
<td>-</td>
<td>59.8</td>
<td>33.6</td>
<td>16.1</td>
<td>56.7</td>
<td>31.1</td>
</tr>
<tr>
<td>VGG16</td>
<td>63.7 (+3.9)</td>
<td>37.2 (+3.6)</td>
<td>17.3 (+1.2)</td>
<td>58.6 (+2.3)</td>
<td>34.0 (+2.9)</td>
<td></td>
</tr>
<tr>
<td>VGG16 138M / 283ms</td>
<td>-</td>
<td>70.4</td>
<td>45.1</td>
<td>24.2</td>
<td>59.2</td>
<td>35.6</td>
</tr>
</tbody>
</table>

- means no distillation or, in other words, trained from scratch.
## Experiments

<table>
<thead>
<tr>
<th>Student</th>
<th>Model Info</th>
<th>Teacher</th>
<th>PASCAL</th>
<th>COCO@.5</th>
<th>COCO@[.5,.95]</th>
<th>KITTI</th>
<th>ILSVRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucker</td>
<td>11M / 47ms</td>
<td>- AlexNet</td>
<td>54.7</td>
<td>25.4</td>
<td>11.8</td>
<td>49.3</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGGM</td>
<td>57.6 (+2.9)</td>
<td>26.5 (+1.2)</td>
<td>12.3 (+0.5)</td>
<td>51.4 (+2.1)</td>
<td>23.6 (+1.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>58.2 (+3.5)</td>
<td>26.4 (+1.1)</td>
<td>12.2 (+0.4)</td>
<td>51.4 (+2.1)</td>
<td>23.9 (+1.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>59.4 (+4.7)</td>
<td>28.3 (+2.9)</td>
<td>12.6 (+0.8)</td>
<td>53.7 (+4.4)</td>
<td>24.4 (+2.1)</td>
</tr>
<tr>
<td>AlexNet</td>
<td>62M / 74ms</td>
<td>- VGGM</td>
<td>57.2</td>
<td>32.5</td>
<td>15.8</td>
<td>55.1</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>59.2 (+2.0)</td>
<td>33.4 (+0.9)</td>
<td>16.0 (+0.2)</td>
<td>56.3 (+1.2)</td>
<td>28.7 (+1.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>60.1 (+2.9)</td>
<td>35.8 (+3.3)</td>
<td>16.9 (+1.1)</td>
<td>58.3 (+3.2)</td>
<td>30.1 (+2.8)</td>
</tr>
<tr>
<td>VGGM</td>
<td>80M / 86ms</td>
<td>- VGG16</td>
<td>59.8</td>
<td>33.6</td>
<td>16.1</td>
<td>56.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>63.7 (+3.9)</td>
<td>37.2 (+3.6)</td>
<td>17.3 (+1.2)</td>
<td>58.6 (+2.3)</td>
<td>34.0 (+2.9)</td>
</tr>
<tr>
<td>VGG16</td>
<td>138M / 283ms</td>
<td>-</td>
<td>70.4</td>
<td>45.1</td>
<td>24.2</td>
<td>59.2</td>
<td>35.6</td>
</tr>
</tbody>
</table>

- means no distillation or, in other words, trained from scratch

Larger teachers lead to greater performance improvements for distilled models

# Experiments

## mAP scores for 5 datasets

<table>
<thead>
<tr>
<th>Student</th>
<th>Model Info</th>
<th>Teacher</th>
<th>PASCAL</th>
<th>COCO@ .5</th>
<th>COCO@[.5,.95]</th>
<th>KITTI</th>
<th>ILSVRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucker</td>
<td>11M / 47ms</td>
<td>- AlexNet</td>
<td>54.7</td>
<td>25.4</td>
<td>11.8</td>
<td>49.3</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGGM</td>
<td>57.6 (+2.9)</td>
<td>26.5 (+1.2)</td>
<td>12.3 (+0.5)</td>
<td>51.4 (+2.1)</td>
<td>23.6 (+1.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>58.2 (+3.5)</td>
<td>26.4 (+1.1)</td>
<td>12.2 (+0.4)</td>
<td>51.4 (+2.1)</td>
<td>23.9 (+1.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>59.4 (+4.7)</td>
<td>28.3 (+2.9)</td>
<td>12.6 (+0.8)</td>
<td>53.7 (+4.4)</td>
<td>24.4 (+2.1)</td>
</tr>
<tr>
<td>AlexNet</td>
<td>62M / 74ms</td>
<td>- VGGM</td>
<td>57.2</td>
<td>32.5</td>
<td>15.8</td>
<td>55.1</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>59.2 (+2.0)</td>
<td>33.4 (+0.9)</td>
<td>16.0 (+0.2)</td>
<td>56.3 (+1.2)</td>
<td>28.7 (+1.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>60.1 (+2.9)</td>
<td>35.8 (+3.3)</td>
<td>16.9 (+1.1)</td>
<td>58.3 (+3.2)</td>
<td>30.1 (+2.8)</td>
</tr>
<tr>
<td>VGGM</td>
<td>80M / 86ms</td>
<td>- VGG16</td>
<td>59.8</td>
<td>33.6</td>
<td>16.1</td>
<td>56.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>63.7 (+3.9)</td>
<td>37.2 (+3.6)</td>
<td>17.3 (+1.2)</td>
<td>58.6 (+2.3)</td>
<td>34.0 (+2.9)</td>
</tr>
<tr>
<td>VGG16</td>
<td>138M / 283ms</td>
<td>-</td>
<td>70.4</td>
<td>45.1</td>
<td>24.2</td>
<td>59.2</td>
<td>35.6</td>
</tr>
</tbody>
</table>

- means no distillation or, in other words, trained from scratch

**Why do you think there are performance improvements from model compression?**

## Experiments

### mAP scores for 5 datasets

<table>
<thead>
<tr>
<th>Student</th>
<th>Model Info</th>
<th>Teacher</th>
<th>PASCAL</th>
<th>COCO@.5</th>
<th>COCO@[.5,.95]</th>
<th>KITTI</th>
<th>ILSVRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tucker</td>
<td>11M / 47ms</td>
<td>- AlexNet</td>
<td>57.6 (+2.9)</td>
<td>26.5 (+1.2)</td>
<td>12.3 (+0.5)</td>
<td>51.4 (+2.1)</td>
<td>23.6 (+1.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGGM</td>
<td>58.2 (+3.5)</td>
<td>26.4 (+1.1)</td>
<td>12.2 (+0.4)</td>
<td>51.4 (+2.1)</td>
<td>23.9 (+1.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>59.4 (+4.7)</td>
<td>28.3 (+2.9)</td>
<td>12.6 (+0.8)</td>
<td>53.7 (+4.4)</td>
<td>24.4 (+2.1)</td>
</tr>
<tr>
<td>AlexNet</td>
<td>62M / 74ms</td>
<td>- VGGM</td>
<td>57.2</td>
<td>32.5</td>
<td>15.8</td>
<td>55.1</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VGG16</td>
<td>59.2 (+2.0)</td>
<td>33.4 (+0.9)</td>
<td>16.0 (+0.2)</td>
<td>56.3 (+1.2)</td>
<td>28.7 (+1.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.1 (+2.9)</td>
<td>35.8 (+3.3)</td>
<td>16.9 (+1.1)</td>
<td>58.3 (+3.2)</td>
<td>30.1 (+2.8)</td>
</tr>
<tr>
<td>VGGM</td>
<td>80M / 86ms</td>
<td>- VGG16</td>
<td>59.8</td>
<td>33.6</td>
<td>16.1</td>
<td>56.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>63.7 (+3.9)</td>
<td>37.2 (+3.6)</td>
<td>17.3 (+1.2)</td>
<td>58.6 (+2.3)</td>
<td>34.0 (+2.9)</td>
</tr>
<tr>
<td>VGG16</td>
<td>138M / 283ms</td>
<td>-</td>
<td>70.4</td>
<td>45.1</td>
<td>24.2</td>
<td>59.2</td>
<td>35.6</td>
</tr>
</tbody>
</table>

- means no distillation or, in other words, trained from scratch

Still, larger models with more parameters return the best results.

Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
Distilling Knowledge for Generative LLMs; Output is an Intractable Set of Many Possible Words; e.g.,

e.g., distill only sampled commonsense knowledge as a tractable symbolic set (and creates human-readable knowledge graph)

Localized Symbolic Knowledge Distillation for Visual Commonsense Models

Jae Sung Park¹, Jack Hessel², Khyathi Raghavi Chandu², Paul Pu Liang²,⁴, Ximing Lu¹,², Peter West¹,², Youngjae Yu⁵, Qiuyuan Huang³, Jianfeng Gao³, Ali Farhadi¹,², Yejin Choi¹,²
One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation

Zhiwei Hao¹,², Jianyuan Guo³, Kai Han², Yehui Tang², Han Hu¹*, Yunhe Wang²*, Chang Xu³
What Knowledge Gets Distilled in Knowledge Distillation?

Utkarsh Ojha*  Yuheng Li*  Anirudh Sundara Rajan*

Yingyu Liang  Yong Jae Lee
Today’s Topics

• Motivation

• Key idea: knowledge distillation (KD)

• Pioneering KD model for image classification

• Pioneering KD model for object detection

• State-of-the-art for KD
The End