Sequential Knowledge Transfer in Teacher-Student Framework Using Densely Distilled Flow-Based Information

Electronics and Telecommunications Research Institute
ICIP 2018
Doyeob Yeo, Ji-Hoon Bae, Junho Yim, Nae-Soo Kim, Cheol-Sig Pyo, and Junmo Kim
Introduction

• Knowledge Distillation

Deep Neural Network

Extract the useful Knowledge

Motivation

• Knowledge Distillation with Teacher – Student framework
Motivation

• Two main issues in knowledge distillation method

What kind of Knowledge?  How to transfer the knowledge?
Previous Research
Previous Research  Knowledge Distillation

  - Student DNN is penalized according to a softened version of the teacher DNN’s output

\[ q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \]

Softened softmax
Previous Research  Knowledge Distillation

  - Student DNN is also penalized according to intermediate features of the teacher DNN
Previous Research Knowledge Distillation

- Determine the distilled knowledge as the flow of the solving procedure calculated with the proposed FSP matrix

\[
G_{i,j}(x; W) = \sum_{s=1}^{h} \sum_{t=1}^{w} \frac{F_{s,t,i}(x; W) \times F_{s,t,j}(x; W)}{h \times w}
\]
Proposed Model
Motivation

• Two main issues in knowledge distillation method

What kind of Knowledge? vs How to transfer the knowledge?
Proposed Model  Sequential Knowledge Transfer

• Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives

http://www.mypearsonstore.com/bookstore/thomas-calculus-9780321587992
Proposed Model  Sequential Knowledge Transfer

• Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
   2.1 Rates of Change and Tangents to Curves
   2.2 Limit of a Function and Limit Laws
   2.3 The Precise Definition of a Limit
   2.4 One-Sided Limits
   2.5 Continuity
   2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
   3.1 Tangents and the Derivative at a Point
   3.2 The Derivative as a Function
   3.3 Differentiation Rules
   3.4 The Derivative as a Rate of Change
   3.5 Derivatives of Trigonometric Functions
   3.6 The Chain Rule
   3.7 Implicit Differentiation
   3.8 Related Rates
   3.9 Linearization and Differentials

4. Applications of Derivatives
   4.1 Extreme Values of Functions
   4.2 The Mean Value Theorem
   4.3 Monotonic Functions and the First Derivative Test
   4.4 Concavity and Curve Sketching
   4.5 Applied Optimization
   4.6 Newton’s Method
   4.7 Antiderivatives
Proposed Model
Sequential Knowledge Transfer

• Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives
Proposed Model
Sequential Knowledge Transfer

- Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives
Proposed Model  Sequential Knowledge Transfer

• Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
  2.1 Rates of Change and Tangents to Curves
  2.2 Limit of a Function and Limit Laws
  2.3 The Precise Definition of a Limit
  2.4 One-Sided Limits
  2.5 Continuity
  2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
  3.1 Tangents and the Derivative at a Point
  3.2 The Derivative as a Function
  3.3 Differentiation Rules
  3.4 The Derivative as a Rate of Change
  3.5 Derivatives of Trigonometric Functions
  3.6 The Chain Rule
  3.7 Implicit Differentiation
  3.8 Related Rates
  3.9 Linearization and Differentials

4. Applications of Derivatives
  4.1 Extreme Values of Functions
  4.2 The Mean Value Theorem
  4.3 Monotonic Functions and the First Derivative Test
  4.4 Concavity and Curve Sketching
  4.5 Applied Optimization
  4.6 Newton's Method
  4.7 Antiderivatives
2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives
Proposed Model  Sequential Knowledge Transfer

• Part of Table of contents of Thomas’ Calculus

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton's Method
4.7 Antiderivatives
Proposed Model
Sequential Knowledge Transfer

• Step 1

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity, Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton’s Method
4.7 Antiderivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity, Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton’s Method
4.7 Antiderivatives

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity, Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton’s Method
4.7 Antiderivatives
Proposed Model
Sequential Knowledge Transfer

• Step 1

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives

(b) \[ \hat{W}_{S_t} = \arg \min_{W_{S_t}} \sum_{i=1}^{n_0} \lambda_i \left\| G_i^T(x, W_{S_t}) - G_i^S(x, W_{S_t}) \right\|_2^2, \]
Proposed Model
Sequential Knowledge Transfer

• Step 2

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

4. Applications of Derivatives
4.1 Extreme Values of Functions
4.2 The Mean Value Theorem
4.3 Monotonic Functions and the First Derivative Test
4.4 Concavity and Curve Sketching
4.5 Applied Optimization
4.6 Newton’s Method
4.7 Antiderivatives
Proposed Model
Sequential Knowledge Transfer

• Step 2

2. Limits and Continuity
2.1 Rates of Change and Tangents to Curves
2.2 Limit of a Function and Limit Laws
2.3 The Precise Definition of a Limit
2.4 One-Sided Limits
2.5 Continuity
2.6 Limits Involving Infinity; Asymptotes of Graphs

3. Differentiation
3.1 Tangents and the Derivative at a Point
3.2 The Derivative as a Function
3.3 Differentiation Rules
3.4 The Derivative as a Rate of Change
3.5 Derivatives of Trigonometric Functions
3.6 The Chain Rule
3.7 Implicit Differentiation
3.8 Related Rates
3.9 Linearization and Differentials

\( (\hat{W}_{S_2}) = \arg \min_{W_{S_2}} \sum_{l=1}^{2n-1} \alpha_l \left\| G_i^T(x, W_{T}) - G_i^S(x, W_{S_2}) \right\|^2_{2}, \quad (2) \)
Proposed Model

Sequential Knowledge Transfer

• Step 3
Proposed Model
Sequential Knowledge Transfer

• Step 3
Proposed Model  Sequential Knowledge Transfer

• Whole procedure

Step 1

Step 2

Step 3
Experimental Result
Experimental Result

- Contents
  - Performance Improvement
  - Network Minimization
Experimental Result

Performance improvement

- Experimental setting
  - CIFAR-10 dataset
  - Using 26-layers Residual network for the Teacher DNN
  - Using 8-layers Residual network for the Student DNN
Experimental Result

Performance improvement

• Experimental setting
  • CIFAR-10, Using 8-layers Residual network
  • 21k Iter for stage 1 and 64k Iter for stage 2
  • 26-layer teacher ResNet with an accuracy of 91.91%

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy [%]</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hint-based method [25]</td>
<td>88.4</td>
<td>$P_c = 87.94%$ for the original eight-layer ResNet</td>
</tr>
<tr>
<td>Original flow-based method [27]</td>
<td>88.72</td>
<td></td>
</tr>
<tr>
<td>Proposed method</td>
<td><strong>88.96</strong></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Result  Performance improvement

- Experimental setting
  - CIFAR-100 dataset
  - Using 32-layers Residual network for the Teacher DNN
  - Using 14-layers Residual network for the Student DNN
Experimental Result

Performance improvement

• Experimental setting
  • CIFAR-100, Using 14-layers Residual network
  • 32k Iter for stage 1 and 64k Iter for stage 2
  • 32-layer teacher ResNet with an accuracy of 64.69%

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy [%]</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hint-based method [25]</td>
<td>63.38</td>
<td>$P_c = 62.37%$ for the original 14-layer ResNet</td>
</tr>
<tr>
<td>Original flow-based method [27]</td>
<td>64.74</td>
<td></td>
</tr>
<tr>
<td>Proposed method</td>
<td><strong>65.06</strong></td>
<td></td>
</tr>
</tbody>
</table>

CIFAR-100
Experimental Result

- Experimental setting
  - CIFAR-10 dataset
  - Using 26-layers Residual network for the Teacher and Student DNN

Teacher DNN
- 64k iterations
- Stage 1

Student DNN
- 21k iterations
- Stage 1
- 21k iterations
- Stage 2
- 21k iterations/each student
Experimental Result  Fast optimization

- Experimental setting
  - CIFAR-10,
  - 21k Iter for stage 1, 21k Iter for stage 2

<table>
<thead>
<tr>
<th>Method</th>
<th>Net1</th>
<th>Net2</th>
<th>Net3</th>
<th>Avg.</th>
<th>Ensemble</th>
<th>#Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original teacher(^a)</td>
<td>91.91</td>
<td>91.68</td>
<td>91.68</td>
<td>91.75</td>
<td>93.29</td>
<td>192k</td>
</tr>
<tr>
<td>Original teacher(^b)</td>
<td>90.92</td>
<td>90.85</td>
<td>90.69</td>
<td>90.82</td>
<td>92.7</td>
<td>63k</td>
</tr>
<tr>
<td>Hint-based method [25]</td>
<td>92.07</td>
<td>91.75</td>
<td>91.81</td>
<td>91.87</td>
<td>93.02</td>
<td>138k</td>
</tr>
<tr>
<td>Original flow-based method [27]</td>
<td>91.84</td>
<td>92.13</td>
<td>92.25</td>
<td>92.07</td>
<td>93.59</td>
<td>126k</td>
</tr>
<tr>
<td>Proposed method</td>
<td>92.36</td>
<td>92.34</td>
<td>92.15</td>
<td>92.28</td>
<td>93.68</td>
<td>126k</td>
</tr>
</tbody>
</table>

\(^a\) The 26-layer teacher ResNet was trained with 64,000 iterations.

\(^b\) The 26-layer teacher ResNet was trained with 21,000 iterations.
Experimental Result

Fast optimization

- Experimental setting
  - Using 32-layers Residual network for the Teacher and Student DNN
- CIFAR-100 dataset

Teacher DNN

- 64k iterations

CIFAR-100 dataset

Student DNN

- 21k iterations
  - Stage 1
  - 21k iterations/each student

- 21k iterations
  - Stage 2

- 21k iterations

Teacher DNN modules include:
- Residual modules 1 (2m)
- Residual modules 2 (2m)
- Residual modules 3 (2m)

Student DNN modules include:
- Residual modules 1 (2m)
- Residual modules 2 (2m)
- Residual modules 3 (2m)

Student DNNs

L_{opt} (W_s)
Experimental Result

Fast optimization

- Experimental setting
  - CIFAR-100,
  - 32k Iter for stage 1, 21k Iter for stage 2

<table>
<thead>
<tr>
<th>Method</th>
<th>Net1</th>
<th>Net2</th>
<th>Net3</th>
<th>Avg.</th>
<th>Ensemble</th>
<th>#Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original teacher&lt;sup&gt;a&lt;/sup&gt;</td>
<td>64.69</td>
<td>63.29</td>
<td>64.52</td>
<td>64.16</td>
<td>69.79</td>
<td>192k</td>
</tr>
<tr>
<td>Original teacher&lt;sup&gt;b&lt;/sup&gt;</td>
<td>62.96</td>
<td>62.69</td>
<td>60.82</td>
<td>62.15</td>
<td>67.91</td>
<td>63k</td>
</tr>
<tr>
<td>Hint-based method [25]</td>
<td>63.54</td>
<td>64.43</td>
<td>64.07</td>
<td>64.01</td>
<td>68.68</td>
<td>168k</td>
</tr>
<tr>
<td>Original flow-based method [27]</td>
<td>64.16</td>
<td>64.3</td>
<td>64.48</td>
<td>64.31</td>
<td>69.5</td>
<td>159k</td>
</tr>
<tr>
<td>Proposed method</td>
<td>66.65</td>
<td>66.52</td>
<td>64.54</td>
<td><strong>65.9</strong></td>
<td><strong>69.98</strong></td>
<td>159k</td>
</tr>
</tbody>
</table>

<sup>a</sup> The 26-layer teacher ResNet was trained with 64,000 iterations.

<sup>b</sup> The 26-layer teacher ResNet was trained with 21,000 iterations.
Conclusion

• Propose a novel approach for enhancing knowledge distillation and knowledge transfer between teacher and student DNN models.

• Help to obtain a fast optimization with high accuracy using the densely distilled flow-based knowledge and its sequential transfer.

• Proposed method outperforms state-of-the-art knowledge transfer method in the network minimization.