Knowledge Distillation\textsuperscript{[1]} (KD)

\[ L_{kd} = H(p^s, p^t) = H(\sigma(a^s; T), \sigma(a^t; T)) = - \sum_{k=1}^{K} p^t[k] \log p^s[k] = -\langle p^t, \log p^s \rangle \]

\[ L_{KD}(W_S) = H(y_{true}, P_S) + \lambda H(P_T^r, P_S^r). \]

- \( T \) is the temperature to soften the logits for more fine-grained information
- \( \sigma(\cdot) \) is the softmax operation with temperature \( T \)
- \( H(\cdot, \cdot) \) is the cross-entropy loss
- \( p[k] \) is the k-th component of vector \( p \)
- \( \langle \cdot, \cdot \rangle \) is the inner product of two vectors (tensors)

Knowledge Distillation (KD)

\[ \mathcal{L}_{kd} = \mathcal{H}(p^s, p^t) = \mathcal{H}(\sigma(a^s; T), \sigma(a^t; T)) = - \sum_{k=1}^{K} p^t[k] \log p^s[k] = -\langle p^t, \log p^s \rangle \]

\[ \mathcal{L}_{KD}(W_S) = \mathcal{H}(y_{\text{true}}, P_S) + \lambda \mathcal{H}(P_T, P_S). \]

Fitnets\[^2\]

\[ \mathcal{L}_{HT}(W_{\text{Guided}}, W_r) = \frac{1}{2} \| u_h(x; W_{\text{Hint}}) - r(v_g(x; W_{\text{Guided}}); W_r) \|^2 \]

- \( u_h \) and \( v_g \) are the teacher/student deep nested functions up to their respective hint/guided layers with parameters \( W_{\text{Hint}} \) and \( W_{\text{Guided}} \)
- \( r \) is the regressor function on top of the guided layer with parameters \( W_r \)

Background

Knowledge Distillation (KD)

$$\mathcal{L}_{kd} = \mathcal{H}(p^s, p^t) = \mathcal{H}(\sigma(a^s; T), \sigma(a^t; T)) = -\sum_{k=1}^{K} p^t[k] \log p^s[k] = -\langle p^t, \log p^s \rangle$$

$$\mathcal{L}_{KD}(W_s) = \mathcal{H}(y_{true}, P_s) + \lambda \mathcal{H}(P^*_T, P^*_S).$$

Fitnets

$$\mathcal{L}_{HT}(W_{Guided}, W_r) = \frac{1}{2}\|u_h(x; W_{Hint}) - r(y(x; W_{Guided}); W_r)\|^2$$

Ensemble KD

$$\mathcal{L}_{ens} = \mathcal{H}(y, \sigma(a^s; 1)) + \lambda \cdot \mathcal{L}_{mkd} + \beta \cdot \mathcal{L}_{mfkd}.$$  

$$\mathcal{L}_{mkd} = \mathcal{H}(p^s, \frac{1}{M} \sum_{m=1}^{M} p^t_m) = -\langle \frac{1}{M} \sum_{m=1}^{M} p^t_m, \log p^s \rangle = \frac{1}{M} \sum_{m=1}^{M} \mathcal{H}(p^s, p^t_m).$$  

$$\mathcal{L}_{mfkd} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{D}(r^t_m(f^t), r^s_m(f^s)).$$
Motivation

Issues:
- averaging neglects the diversity of different teachers
- there might be conflicts, competition or even noise among all teachers
- the final direction will be determined by some dominant teachers
Motivation

Issues:
• averaging neglects the diversity of different teachers
• there might be conflicts, competition or even noise among all teachers
• the final direction will be determined by some dominant teachers

Our Proposed Method:
• formulate ensemble knowledge distillation as a multi-objective optimization problem
• apply multiple-gradient descent algorithm to probe an optimal solution
• the optimizing direction will be less influenced by noisy teachers
• introduce a tolerance parameter to control the disagreement among teachers.
In the training of the student network, gradients provided by teacher networks can serve as learning directions.

The final direction $d$ should accommodate as more teachers as possible.

Obtaining such a proper direction can resort to gradient-based MOO methods\[3\]

\[
\min_{\alpha} \frac{1}{2} \left\| \sum_{m=1}^{M} \alpha_m \nabla \ell_m^t(\theta^{(\tau)}) \right\|^2, \text{s.t. } \sum_{m=1}^{M} \alpha_m = 1, \ 0 \leq \alpha_m \leq C, \ \forall m \in [1 : M]
\]

\[
d^{(\tau)} = -\sum_{m=1}^{M} \alpha_m^* \nabla \ell_m^t(\theta^{(\tau)})
\]

$C \in [1/M, 1]$ acts as a constant controlling the tolerance of disagreement among teachers.

Framework

Figure 1

Figure 2
Framework

Table 1: Classification accuracy (%) of student network on CIFAR10 and CIFAR100 with resnet56 teachers. \# indicates the ensemble size of teachers. "*" refers to feature-based ensemble KD using AE-KD method. \(\uparrow\) means the performance improvement of AE-KD with respect to the baseline AVER. Ens refers to the "Ensemble", and we use majority voting to calculate the performance of ensemble.

<table>
<thead>
<tr>
<th></th>
<th>CIFAR10</th>
<th>MobileNetV2 (75.97)</th>
<th></th>
<th>CIFAR100</th>
<th>MobileNetV2 (64.60)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resnet20 (91.7)</td>
<td></td>
<td>resnet20 (69.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>Ens</td>
<td>AVER</td>
<td>AE-KD</td>
<td>FitNets</td>
<td>AE-KD*</td>
</tr>
<tr>
<td>1</td>
<td>93.94</td>
<td>91.84</td>
<td>91.84</td>
<td>91.86</td>
<td>91.86</td>
</tr>
<tr>
<td>5</td>
<td>95.27</td>
<td>91.94</td>
<td>92.50 ((\uparrow)0.56)</td>
<td>91.96</td>
<td>92.58 ((\uparrow)0.62)</td>
</tr>
<tr>
<td>10</td>
<td>95.58</td>
<td>91.96</td>
<td>92.54 ((\uparrow)0.58)</td>
<td>92.09</td>
<td>92.67 ((\uparrow)0.58)</td>
</tr>
<tr>
<td>15</td>
<td>95.75</td>
<td>92.01</td>
<td>92.55 ((\uparrow)0.54)</td>
<td>92.18</td>
<td>92.89 ((\uparrow)0.71)</td>
</tr>
<tr>
<td>20</td>
<td>95.74</td>
<td>92.22</td>
<td>92.70 ((\uparrow)0.48)</td>
<td>92.31</td>
<td>92.93 ((\uparrow)0.62)</td>
</tr>
<tr>
<td>25</td>
<td>95.78</td>
<td>92.50</td>
<td>92.74 ((\uparrow)0.24)</td>
<td>92.50</td>
<td>93.01 ((\uparrow)0.51)</td>
</tr>
</tbody>
</table>

Table 2: Classification accuracy (%) of student networks with the same architecture as teachers on CIFAR10 and CIFAR100. Symbols are the same with Table 1.

<table>
<thead>
<tr>
<th>#</th>
<th>Ens</th>
<th>AVER</th>
<th>AE-KD</th>
<th>FitNet</th>
<th>AE-KD*</th>
<th>Ens</th>
<th>AVER</th>
<th>AE-KD</th>
<th>FitNet</th>
<th>AE-KD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93.94</td>
<td>92.58</td>
<td>92.58</td>
<td>92.49</td>
<td>92.49</td>
<td>72.41</td>
<td>73.58</td>
<td>73.58</td>
<td>73.61</td>
<td>73.61</td>
</tr>
<tr>
<td>5</td>
<td>95.27</td>
<td>92.67</td>
<td>92.88 (↑0.21)</td>
<td>92.64</td>
<td>93.47 (↑0.83)</td>
<td>77.42</td>
<td>74.39</td>
<td>74.88 (↑0.49)</td>
<td>74.58</td>
<td>75.02 (↑0.44)</td>
</tr>
<tr>
<td>10</td>
<td>95.58</td>
<td>92.82</td>
<td>93.23 (↑0.41)</td>
<td>92.67</td>
<td>93.64 (↑0.97)</td>
<td>79.15</td>
<td>74.66</td>
<td>74.96 (↑0.30)</td>
<td>74.69</td>
<td>75.16 (↑0.47)</td>
</tr>
<tr>
<td>15</td>
<td>95.75</td>
<td>92.89</td>
<td>93.71 (↑0.82)</td>
<td>92.77</td>
<td>93.66 (↑0.89)</td>
<td>79.55</td>
<td>74.74</td>
<td>75.14 (↑0.40)</td>
<td>74.75</td>
<td>75.37 (↑0.62)</td>
</tr>
<tr>
<td>20</td>
<td>95.74</td>
<td>92.87</td>
<td>93.75 (↑0.88)</td>
<td>92.90</td>
<td>93.76 (↑0.86)</td>
<td>79.83</td>
<td>74.92</td>
<td>75.45 (↑0.53)</td>
<td>75.04</td>
<td>75.53 (↑0.49)</td>
</tr>
<tr>
<td>25</td>
<td>95.78</td>
<td>93.73</td>
<td>94.20 (↑0.47)</td>
<td>93.13</td>
<td>93.88 (↑0.75)</td>
<td>80.01</td>
<td>75.04</td>
<td>75.62 (↑0.58)</td>
<td>75.10</td>
<td>75.69 (↑0.59)</td>
</tr>
</tbody>
</table>

Table 3: Accuracy (%) on ImageNet of ResNet18 student and ResNet50 teacher networks. Symbols are the same with Table 1. [6]

<table>
<thead>
<tr>
<th>#</th>
<th>Ensemble</th>
<th>AVER</th>
<th>AE-KD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
</tr>
<tr>
<td>1</td>
<td>75.67</td>
<td>92.50</td>
<td>67.81</td>
</tr>
<tr>
<td>3</td>
<td>76.85</td>
<td>93.60</td>
<td>67.85</td>
</tr>
<tr>
<td>5</td>
<td>77.52</td>
<td>93.85</td>
<td>68.18</td>
</tr>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
</tr>
<tr>
<td>1</td>
<td>67.81</td>
<td>88.21</td>
<td>67.81</td>
</tr>
<tr>
<td>3</td>
<td>67.85</td>
<td>88.39</td>
<td>68.28 (↑0.43)</td>
</tr>
<tr>
<td>5</td>
<td>69.14 (↑0.96)</td>
<td>88.93 (↑0.46)</td>
<td></td>
</tr>
</tbody>
</table>

Ablation Study

- Tolerance for disagreement:
  - 0.1: 91.4
  - 0.2: 91.6
  - 0.3: 91.8
  - 0.4: 92
  - 0.5: 92.2
  - 0.6: 92.4
  - 0.7: 92.6
  - 0.8: 92.8
  - 0.9: 93
  - 1: 93

- Accuracy (%):
  - Ensemble Size: 10
  - Ensemble Size: 20

- Graphs:
  - Left: Accuracy (%) vs Ensemble Size
  - Right: Accuracy (%) vs Tolerance for disagreement
Thanks for listening