On the Efficacy of Knowledge Distillation - Supplementary Materials

Jang Hyun Cho
Cornell University
jc2926@cornell.edu

Bharath Hariharan
Cornell University
bh497@cornell.edu

Student: WRN16-1				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
-	-	60/200	-	8.751 ± 0.129
DN40-24	0.69 M	60/200	5.419	8.350 ± 0.195
DN100-12	0.77 M		4.974	8.297 ± 0.069
DN40-48	2.73 M		4.667	8.370 ± 0.212
DN100-24	3.02 M		4.272	8.763 ± 0.178
DN40-24	0.69 M	15/50	6.823	8.045 ± 0.092
DN100-12	0.77 M		6.615	7.915 ± 0.120
DN40-48	2.73 M		5.666	**7.854 ± 0.127**
DN100-24	3.02 M		5.435	8.016 ± 0.223

Table 1. WideResNet16-1 trained with different DenseNet teachers. First number next to "DN" indicates depth, followed by growth factor (consistent with the original paper). Top-row shows the result of WideResNet16-1 trained from scratch. In all cases, student trained with early-stopped DeseNet teacher performs better by large margin.

1. More Results on CIFAR10

Here we report more results and details of experiments in our work. Consistent with the main paper, "WRN" and "DN" stand for WideResNet and DenseNet, respectively. Table 7 and 8 show the efficacy of early-stopped teachers for student network WideResNet16-1 and WideResNet28-1 trained from teachers with varying width factor. As stated in the main paper, the number of total epochs $N \in \{35, 50, 65, 80, 200\}$ and learning rate decay step size $k \in \{10, 15, 20, 25, 60\}$ were considered in this experiment. Table 9 shows that our conclusions are consistent with different knowledge distillation method such as attention transfer ("AT+KD"). Table 1, 2, 3, and 6 show different experiment settings (different student-teacher pairs, learning method, etc.)

2. Details on ImageNet Experiments

Here we report more details of ImageNet experiments. Figure 1 are comparisons of different student accuracy plots, showing the harming effect of distillation. Table 4 shows the fully-trained and early-stopped models used as a teacher for ImageNet experiments.

Student: DN40-12				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
-	-	60/200	-	7.268 ± 0.148
DN40-12	0.18 M	60/200	7.169	6.821 ± 0.226
DN40-24	0.69 M	60/200	5.419	6.964 ± 0.139
DN100-12	0.77 M		4.974	6.847 ± 0.278
DN40-48	2.73 M		4.667	7.266 ± 0.359 *
DN100-24	3.02 M		4.272	7.507 ± 0.204 *
DN40-24	0.69 M	15/50	6.823	6.981 ± 0.112
DN100-12	0.77 M		6.615	**6.645 ± 0.089**
DN40-48	2.73 M		5.666	6.679 ± 0.123
DN100-24	3.02 M		5.435	6.721 ± 0.298

Table 2. DenseNet40-12 trained with different DenseNet teachers. First number next to "DN" indicates depth, followed by growth factor (consistent with the original paper). Top-row shows the result of DenseNet40-12 trained from scratch. In all cases student trained with early-stopped DeseNet teacher performs better by large margin. Numbers with * indicate that the students failed to achieve the same accuracy of student trained from scratch.

Student	Teacher	Schedule Type	Error (%)
WRN16-1	WRN16-8	Cosine	7.945 ± 0.127
	WRN16-8 (20/65)	Cosine	**7.781 ± 0.201**
	WRN100-1	Cosine	8.524 ± 0.182
	WRN100-1 (20/65)	Cosine	**8.191 ± 0.104**

Table 3. CIFAR10 results of knowledge distillation with a different learning rate decaying schedule, "Cosine" scheduling. Student trained with early-stopped teacher performed better.

Model	# params	Top 1 Error (%)	Top 5 Error (%)
ResNet18	11.69 M	30.24	10.92
ResNet34	21.79 M	26.70	8.58
ResNet34 (50)	21.79 M	27.72	9.10
ResNet50	25.56 M	23.85	7.13
ResNet50 (35)	25.56 M	27.01	8.75
ResNet152	60.19 M	21.69	6.03
ResNet152 (35)	60.19 M	23.58	7.03

Table 4. Details of models trained from scratch that are used as teachers for ImageNet experiments in the main paper. Models with a number inside parentheses are early-stopped.

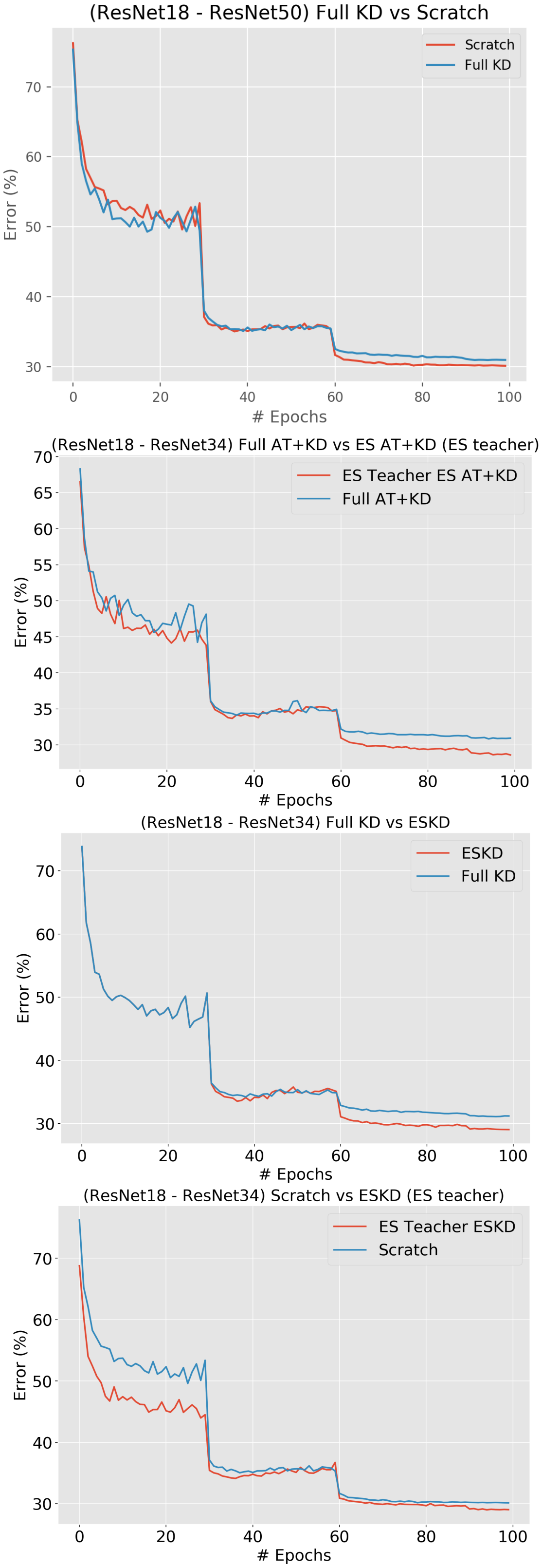

Figure 1. Different plots showing the harming effect of knowledge distillation when student capacity is limited, and how early-stopping mitigates the effect.

Student: WRN16-1				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
-	-	60/200	-	8.751 ± 0.129
WRN40-1	0.56M	60/200	6.517	8.324 ± 0.111
WRN52-1	0.76 M		6.042	8.481 ± 0.198
WRN64-1	0.95 M		6.032	8.573 ± 0.158
WRN76-1	1.15 M		5.864	8.666 ± 0.121
WRN88-1	1.34 M		5.686	8.811 ± 0.153
WRN100-1	1.54 M		5.568	8.484 ± 0.182
WRN154-1	2.41 M		5.478	8.546 ± 0.181
WRN250-1	3.97 M		5.271	8.787 ± 0.173
WRN100-1	1.54 M	15/50	7.526	**8.192 ± 0.198**
WRN154-1	2.41 M		7.318	8.227 ± 0.212
WRN250-1	3.97 M		6.893	8.263 ± 0.234

Table 5. WideResNet16-1 trained with teachers varying depth factor. All students trained with early-stopped teacher performed better than any of students trained from fully-trained teacher by large margin. Among ones with fully-trained teachers, larger models did not make better student. All results are consistent with our conclusions. Note that WideResNet with width factor 1 is equivalent to Pre-Activated ResNet.

Student: DN40-12				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
-	-	60/200	-	7.268 ± 0.148
WRN40-1	0.56M	60/200	6.517	7.389 ± 0.244
WRN52-1	0.76 M		6.042	7.640 ± 0.204
WRN64-1	0.95 M		6.032	7.600 ± 0.247
WRN76-1	1.15 M		5.864	7.407 ± 0.137
WRN88-1	1.34 M		5.686	7.642 ± 0.131
WRN100-1	1.54 M		5.568	7.693 ± 0.134
WRN154-1	2.41 M		5.478	7.780 ± 0.299
WRN250-1	3.97 M		5.271	7.711 ± 0.152
WRN100-1	1.54 M	15/50	7.526	**7.025 ± 0.182**
WRN154-1	2.41 M		7.318	7.169 ± 0.161
WRN250-1	3.97 M		6.893	7.488 ± 0.291

Table 6. DenseNet40-12 trained with WideResNet teachers varying depth factor. All students trained with early-stopped teacher performed better than any of students trained from fully-trained teacher by large margin. Among ones with fully-trained teachers, larger models did not make better student. All results are consistent with our conclusions. Note that WideResNet with width factor 1 is equivalent to Pre-Activated ResNet.

Student: WRN16-1				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
WRN16-1	0.17M	60/200	8.751	8.182 ± 0.250
WRN16-2	0.69M	60/200	6.269	7.610 ± 0.222
WRN16-3	1.55M	60/200	5.340	7.681 ± 0.259
		25/80	6.289	7.517 ± 0.212
		20/65	6.507	**7.498 ± 0.201**
		15/50	6.734	7.788 ± 0.112
		10/35	7.416	8.093 ± 0.119
WRN16-4	2.74M	60/200	4.964	7.733 ± 0.186
		25/80	5.666	7.658 ± 0.062
		20/65	5.963	**7.612 ± 0.112**
		15/50	6.358	7.788 ± 0.112
		10/35	7.130	8.093 ± 0.119
WRN16-6	6.17M	60/200	4.529	7.929 ± 0.071
		25/80	5.261	7.687 ± 0.157
		20/65	5.498	**7.594 ± 0.173**
		15/50	5.893	7.685 ± 0.163
		10/35	6.635	7.751 ± 0.157
WRN16-8	10.96M	60/200	4.410	8.028 ± 0.136
		25/80	4.984	7.642 ± 0.163
		20/65	5.270	**7.482 ± 0.041**
		15/50	5.498	7.596 ± 0.089
		10/35	6.240	7.784 ± 0.223

Table 7. WideResNet16-1 trained with different teachers, and each teacher we performed different “shrinking” of the learning schedule. Step size $k \in \{10, 15, 20, 25, 60\}$ and total number of epoch $N \in \{35, 50, 65, 80, 200\}$ were considered. For kinds of teacher network, students trained with any of the early-stopped teachers outperforms the model trained with fully-trained teacher.

Student: WRN28-1				
Teacher	# params	Mode	Teacher Error (%)	Student Error (%)
WRN28-1	0.36M	60/200	7.101	7.101± 0.072
WRN28-2	1.46M	60/200	5.201	6.973 ± 0.130
WRN28-3	3.29M	60/200	4.687	6.952 ± 0.138
		25/80	5.369	6.702 ± 0.159
		20/65	5.696	**6.621 ± 0.066**
		15/50	6.180	6.544 ± 0.262
		10/35	6.962	6.807 ± 0.076
WRN28-4	5.84M	60/200	4.509	7.118 ± 0.198
		25/80	4.994	6.768 ± 0.099
		20/65	5.201	6.772 ± 0.060
		15/50	5.824	**6.610 ± 0.330**
		10/35	6.526	6.718 ± 0.063
WRN28-6	13.14M	60/200	4.104	7.070 ± 0.159
		25/80	4.608	6.869± 0.152
		20/65	4.865	6.920 ± 0.114
		15/50	5.330	6.720 ± 0.060
		10/35	5.992	**6.710 ± 0.241**
WRN28-8	23.25M	60/200	4.064	7.227 ± 0.149
		25/80	4.578	6.819 ± 0.155
		20/65	4.657	6.817± 0.117
		15/50	5.092	**6.748 ± 0.118**
		10/35	6.022	6.795 ± 0.123

Table 8. WideResNet28-1 trained with different teachers, and each teacher we performed different “shrinking” of the learning schedule. Step size $k \in \{10, 15, 20, 25, 60\}$ and total number of epoch $N \in \{35, 50, 65, 80, 200\}$ were considered. For kinds of teacher network, students trained with any of the early-stopped teachers outperforms the model trained with fully-trained teacher.

Student	# params	Error (%)	Teacher	# params	Teacher Error (%)	Method	Student Error (%)
WRN16-1	0.17M	8.751	WRN16-1	0.17M	8.751	KD	8.182 ± 0.250
			WRN16-2	0.69M	6.269	KD	**7.610 ± 0.222**
			WRN16-3	1.55M	5.340	KD	7.681 ± 0.259
			WRN16-4	2.74M	4.964	KD	7.733 ± 0.186
			WRN16-6	6.17M	4.529	KD	7.929 ± 0.071
			WRN16-8	10.96M	4.410	KD	8.028 ± 0.136
WRN16-1	0.17M	8.751	WRN16-2	0.69M	6.269	AT+KD	**7.498 ± 0.062**
			WRN16-3	1.55M	5.340	AT+KD	7.551 ± 0.130
			WRN16-4	2.74M	4.964	AT+KD	7.656 ± 0.131
			WRN16-6	6.17M	4.529	AT+KD	7.668 ± 0.139
			WRN16-8	10.96M	4.410	AT+KD	7.794 ± 0.203
WRN16-1	0.17M	8.751	WRN16-3 (20/65)	1.55M	6.507	AT+KD	7.498 ± 0.201
			WRN16-4 (20/65)	2.74M	5.963	AT+KD	7.585 ± 0.165
			WRN16-6 (20/65)	6.17M	5.498	AT+KD	**7.484 ± 0.223**
			WRN16-8 (20/65)	10.96M	5.270	AT+KD	7.494 ± 0.165
WRN28-1	0.36M	7.101	WRN28-1	0.36M	7.101	KD	7.101 ± 0.072
			WRN28-2	1.46M	5.201	KD	6.973 ± 0.130
			WRN28-3	3.29M	4.687	KD	**6.952 ± 0.138**
			WRN28-4	5.84M	4.509	KD	7.118 ± 0.198
			WRN28-6	13.14M	4.104	KD	7.070 ± 0.159
			WRN28-8	23.35M	4.064	KD	7.227 ± 0.149
WRN28-1	0.36M	7.101	WRN28-2	1.46M	5.201	AT+KD	6.538 ± 0.185
			WRN28-3	3.29M	4.687	AT+KD	6.526 ± 0.121
			WRN28-4	5.84M	4.509	AT+KD	6.657 ± 0.118
			WRN28-6	13.14M	4.104	AT+KD	**6.443 ± 0.092**
			WRN28-8	23.35M	4.064	AT+KD	6.487 ± 0.222
WRN28-1	0.36M	7.101	WRN28-3 (15/50)	3.29M	6.180	AT+KD	6.410 ± 0.162
			WRN28-4 (15/50)	5.84M	5.824	AT+KD	6.429 ± 0.090
			WRN28-6 (15/50)	13.14M	5.330	AT+KD	**6.358 ± 0.168**
			WRN28-8 (15/50)	23.35M	5.092	AT+KD	6.402 ± 0.107

Table 9. WideResNet16-1 and WideResNet28-1 trained with teachers of increasing width. Attention transfer method was also explored. Teachers with (k/N) indicate early-stopped (step size/ total epochs). Our conclusions are consistent with different method such as attention transfer.

Student	# params	Error (%)	Teacher	# params	Teacher Error (%)	Method	Student Error (%)
WRN28-1	0.36M	7.101	WRN16-4	2.74M	4.964	KD	6.518 ± 0.204
			WRN16-4 (20/65)		5.963	KD	**6.483 ± 0.173**
			WRN16-4		4.964	AT+KD	6.357 ± 0.086
			WRN16-4 (20/65)		5.963	AT+KD	**6.253 ± 0.177**
WRN28-1	0.36M	7.101	WRN16-6	6.17M	4.529	KD	6.613 ± 0.227
			WRN16-6 (20/65)		5.498	KD	**6.230 ± 0.069**
			WRN16-6		4.529	AT+KD	6.253 ± 0.278
			WRN16-6 (20/65)		5.498	AT+KD	**6.133 ± 0.113**
WRN28-1	0.36M	7.101	WRN16-11	20.70M	4.193	KD	6.774 ± 0.111
			WRN16-11 (20/65)		5.033	KD	**6.281 ± 0.184**
			WRN16-11		4.193	AT+KD	6.360 ± 0.109
			WRN16-11 (20/65)		5.033	AT+KD	**6.202 ± 0.172**

Table 10. WideResNet28-1 student trained with even shallower teachers (WideResNet16-x) on CIFAR10. Consistent with our conclusions, early-stopped teachers produce better student. The teachers are chosen to be compared to WRN28-3, WRN28-4, and WRN28-8 in terms of the number of the parameters.