Knowledge Condensation Distillation (Supplementary Material)

Chenxin Li[1], Mingbao Lin[2], Zhiyuan Ding[1], Nie Lin[4], Yihong Zhuang[1],
Yue Huang[1,3⋆], Xinghao Ding[1,3], and Liujuan Cao[1]

[1] School of Informatics, Xiamen University
[2] Tencent Youtu Lab
[3] Institute of Artificial Intelligence, Xiamen University
[4] Hunan University
chenxinli@stu.xmu.edu.cn linmb001@outlook.com
dingzhiyuan@stu.xmu.edu.cn nielin@hnu.edu.cn
{zhuangyihong,yhuang2010,dxh,caoliujuan}@xmu.edu.cn

A Additional Experimental Results

A.1 Calculation of Computation Cost

In what follows, we first shortly describe the metrics of computation cost in UNIX [1] and then show that they are equivalent to the metrics of cost C used in this paper (Sec. 3.5).

Computation Metrics in UNIX. Considering the sampling number and the computation in different network passes, the cost of computation in UNIX is calculated by:

$$E = N_t \cdot F_t + N_{s1} \cdot F_s + N_{s2} \cdot B_s, \tag{17}$$

where F_t, F_s and B_s denote the float-point operation number in teacher forward pass, student forward pass and student backward pass while N_t, N_{s1}, N_{s2} denote their total sampling number over the entire training procedure.

For a vanilla KD baseline, the sampling number in different passes keeps fixed, *i.e.*, $N_t = N_{s1} = N_{s2}$. By denoting this value as N, the baseline cost can be derived as: $E = N \cdot (F_t + F_s + B_s)$. As a comparison, for UNIX, N_t and N_{s2} are reduced to N_k (where $N_k < N$) while N_{s1} is increased to $N + N_k$, which makes $E = N_k \cdot F_t + (N + N_k) \cdot F_s + N_k \cdot B_s$. *Computation* is calculated by the ratio of them:

$$\begin{aligned} Computation &= \frac{N_k \cdot F_t + (N + N_k) \cdot F_s + N_k \cdot B_s}{N \cdot F_t + N \cdot F_s + N \cdot B_s} \\ &= \frac{N_k \cdot 1 + (N + N_k) \cdot \frac{F_s}{F_t} + N_k \cdot \frac{B_s}{F_t}}{N \cdot 1 + N \cdot \frac{F_s}{F_t} + N \cdot \frac{B_s}{F_t}}. \end{aligned} \tag{18}$$

In UNIX [1], the approximation of $B_s \approx F_s$ is introduced, so Eq. (18) can be re-written as:

$$Computation \approx \frac{N_k \cdot 1 + (N + 2N_k) \cdot \frac{F_s}{F_t}}{N \cdot 1 + 2N \cdot \frac{F_s}{F_t}}. \tag{19}$$

[⋆] Corresponding Author

where $\frac{F_s}{F_t}$ denotes the ratio of float-point operation number between student and teacher forward passes, which varies in different teacher-student pairs.

Relation of Computation Metrics Between This Paper and UNIX. In this paper, the absolute computation cost C_a counts the total sampling number over the training procedure. Formally, given I training epochs, T epochs in a condensation stage, condensation thresholdτ^t at the t-th stage $C_a = |K| \cdot I$ provides the computation for the vanilla KD baseline. For our KCD, $C_a = |K| \cdot (\tau^0 + \dots + \tau^t + \dots + \tau^{I/T}) \cdot T$. By calculating the ratio, the relative computation cost C can be obtained, as shown in Eq. (16).

In fact, if we analyze the computation of our KCD by Eq. (17), we have:

$$E_{KCD} = |K| \cdot (\tau^0 + \dots + \tau^t + \dots + \tau^{I/T}) \cdot T \cdot (F_t + F_s + B_s). \tag{20}$$

Further considering the vanilla KD baseline as $E = |K| \cdot I \cdot (F_t + F_s + B_s)$, the $Computation$ for our KCD can be calculated as:

$$\begin{aligned} Computation_{KCD} &= \frac{|K| \cdot (\tau^0 + \dots + \tau^t + \dots + \tau^{I/T}) \cdot T \cdot \cancel{(F_t + F_s + B_s)}}{|K| \cdot I \cdot \cancel{(F_t + F_s + B_s)}} \\ &= C. \end{aligned} \tag{21}$$

Thus the metric of *computation* in UNIX is equivalent to C in Eq. (16) of the main paper. This means that the computation results of our KCD are comparable with those in UNIX, as shown in Tab. 2 and Tab. 3 of our main paper. It is noteworthy that the *computation* or C for our KCD is only dependent on the compactness of knowledge encoding, thus keeps unchanged across different teacher-student pairs.

Results in Tab. 2 of Our Main Paper. In Tab. 2 of our main paper, $C = 100\%$ is set for the vanilla KD baseline while $C = 81.6\%$ is calculated via the derivation of Eq. (16) in our main paper, as $C = \frac{\rho^{\frac{I}{T}}(1-\rho)}{1-\rho^{\frac{T}{T}}} \cdot \frac{T}{I}$ (derived from the exponential decaying of the condensation threshold τ). For the second row, we directly cite the results in UNIX with the most similar *computation* or C to ours. For the third row, we run the official code of UNIX with the adjusted parameters of N_k in Eq. (19) to make $Computation$ be equal to our KCD. We can observe that the accuracy of the proposed KCD outperforms UNIX at the same level of computation cost.

Results in Tab. 3 of Our Main Paper. $C = 81.6\%$ on ImageNet is calculated in the same way with CIFAR100 in Tab. 2 of our main paper. C_a of vanilla KD baseline is calculated via $1.26M \times 90 = 114M$. C_a of the proposed KCD is calculated by $114M \times C = 81M$. As can be seen, our KCD reveals an obvious gain of distillation accuracy and efficiency on the large-scale benchmark.

A.2 More Ablation Studies

Influence of Cost-Aware Weighting Coefficient $\boldsymbol{\alpha}$. We evaluate the sensitivity of the proposed KCD *w.r.t.* $\boldsymbol{\alpha}$ (in Eq. (13) of our main paper) in three KD

Table 5: Ablation study of α on CIFAR100 (Acc%).

Teacher Student	wrn-40-2 wrn-16-2	vgg13 mobilenetv2	resnet32x4 ShuffleNetV2
0.01	75.27	66.40	74.85
0.03	**75.70**	**68.61**	75.19
0.05	74.99	67.86	**75.23**
0.1	75.43	68.28	75.11

Table 6: Ablation study of ρ on CIFAR100 (Acc%).

Teacher Student	wrn-40-2 wrn-16-2	vgg13 mobilenetv2	resnet32x4 ShuffleNetV2
0.9	75.01	68.23	**75.36**
0.7	**75.70**	**68.61**	75.19
0.5	74.89	66.81	75.08
0.3	74.07	65.50	74.03

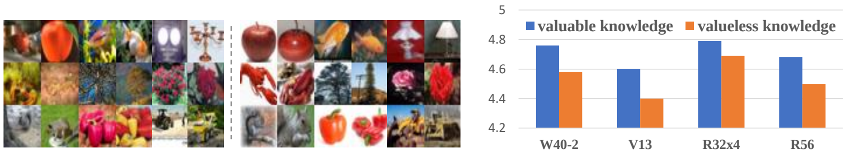
Table 7: Ablation study on CIFAR100 (Acc%).

Teacher Student	wrn-40-2 wrn-16-2	vgg13 mobilenetv2	resnet32x4 ShuffleNetV2
K_1 w/o Aug.	75.45	68.23	75.16
Aug. K_1 in random ($\epsilon = \epsilon_m$)	75.41	65.94	75.12
Aug. K_{1H}	75.52	68.02	74.77
Aug. K_{1L}	**75.70**	**68.61**	**75.19**

processes on CIFAR100. The results are reported in Tab. 5. We vary the parameter in {0.01, 0.03, 0.05, 0.1}, and choose $\boldsymbol{\alpha} = \mathbf{0.03}$ for its best performance.

Influence of Final Condensation Ratio $\boldsymbol{\rho}$. We further evaluate the sensitivity of the proposed KCD *w.r.t.* final condensation ratio of knowledge encoding, $\boldsymbol{\rho}$ (in Alg. 1 of our main paper) on CIFAR100. The results are reported in Tab. 6. We vary the parameter in {0.9, 0.7, 0.5, 0.3}, and choose $\boldsymbol{\rho} = \mathbf{0.7}$ due to its best performance.

Design of Knowledge Augmentation. Further, we evaluate the performance of the proposed value-adaptive knowledge summary (VAKS) module equipped with different knowledge augmentation strategies on CIFAR100. The results are displayed in Tab. 7. **$\boldsymbol{K_1}$ w/o Aug.** denotes that we remove the knowledge augmentation in our work, letting $\hat{K} = K_1$ in our main paper. **Aug. $\boldsymbol{K_1}$ in random** denotes that we utilize K_0 to randomly augment $|K_0|$ knowledge points in K_1 with $\epsilon = \epsilon_m$ (as in Eq. (14) of our main paper). **Aug. $\boldsymbol{K_{1H}}$** denotes that we re-partition K_1 to K_{1H} and K_{1L} so that $|K_{1H}| = |K_0|$, and augment K_{1H}. **Aug. $\boldsymbol{K_{1L}}$** corresponds to the final design in our main paper, which achieves the best performance.



(a) Images of valuable or valueless knowledge (b) Average entropy of soft labels

Fig. 5: Visualization of images and knowledge hints in knowledge points with value label $y = 1$ or 0.

A.3 Visualization of Valuable Knowledge Points

Fig. 5 displays the visualization of the knowledge points with value label $y = 1$ or 0, which is conducted on CIFAR100. As shown in Fig. 5(a)(**Left**), the images from "valueless" knowledge show less complexity where objects are centered and easily distinguishable. In comparison, the "valuable" images shown in Fig. 5(a)(**Right**) tend to contain multiple ambiguous elements that are lower-quality with complicated backgrounds and more challenging to recognize. Fig. 5(b) reveals the "valuable" patterns of knowledge hints from the average entropy of soft labels via four teacher models. It appears that the soft labels of valuable knowledge points tend to have a higher entropy, implying more informative semantic structural information.

References

1. Xu, G., Liu, Z., Loy, C.C.: Computation-efficient knowledge distillation via uncertainty-aware mixup. arXiv preprint arXiv:2012.09413 (2020)