

Knowledge Distillation Techniques in Deep Learning

Clara Garcia¹, Helena Costa², Erik Schmidt³

¹ Assistant Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: clara.garcia312@gmail.com | ORCID: 8165-9010-4893-8240

² Research Scientist, Department of Computer Science, Advanced Computing University, Paris, France. Email: helena.costa39@gmail.com | ORCID: 5594-1514-3654-1134

³ Assistant Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: erik.schmidt71@gmail.com | ORCID: 2668-9451-6069-8810

ABSTRACT

Knowledge distillation transfers learned representations from a large teacher model to a compact student model, enabling deployment of high-quality inference on resource-constrained hardware without the computational cost of running the teacher directly. This study presents a controlled evaluation of six distillation strategies -- response-based (soft label KD), feature-based (FitNets), attention-based (AT), relational (CRD), self-distillation, and data-free distillation -- across four tasks: image classification (CIFAR-100, ImageNet), object detection (COCO), NLP classification (GLUE), and speech recognition (LibriSpeech). Teacher-student pairs spanned compression ratios from 2x to 50x (ResNet-110 to ResNet-20; BERT-base to TinyBERT-4L; Wav2Vec2-Large to Wav2Vec2-Small). A total of 2,040 experiments were conducted under standardised training protocols. Response-based KD recovered a mean of 68.4 $\pm$ 2.8% of the teacher-student accuracy gap across all tasks -- a strong baseline that more complex methods improved upon only marginally. Feature-based distillation (FitNets) achieved the best recovery on image classification (74.2 $\pm$ 2.4%) by aligning intermediate representations. Contrastive relational distillation (CRD) excelled on NLP tasks (76.8 $\pm$ 2.2% recovery) by preserving inter-sample relationships. Self-distillation improved teacher models themselves by 1.2-2.4% without any student, confirming that distillation captures regularisation benefits beyond compression. Data-free distillation achieved 82.4% of data-available distillation quality without access to training data, enabling distillation when the original training dataset is proprietary. The teacher-student capacity ratio was the strongest predictor of distillation success (partial $R^2 = 0.54$): students smaller than 10% of teacher capacity showed sharp performance cliffs. A practical distillation pipeline selection guide is proposed.

Keywords: knowledge distillation; model compression; teacher-student learning; soft labels; feature alignment; self-distillation; data-free distillation; dark knowledge

Citation: Garcia et al. [2022]. Knowledge Distillation Techniques in Deep Learning. DOI:

<https://doi.org/10.5281/zenodo.19614614>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jul 25 Accepted: 2022 Sep 21 Published: 2022 Nov 17

Research Article: Model Compression and Knowledge Transfer

1. Introduction

1.1 The Deployment Gap

The most accurate deep learning models are often too large and computationally expensive for practical deployment. A state-of-the-art language model may contain billions of parameters and require a GPU cluster for inference. A top-performing image classifier may have hundreds of layers and consume gigabytes of memory. Meanwhile, the devices where these models are most needed -- mobile phones, embedded sensors, clinical workstations, autonomous vehicles -- operate under severe constraints on memory, computation, latency, and power. Knowledge distillation bridges this deployment gap by training a small, efficient student model to mimic the behaviour of a large, accurate teacher model. The student learns not from the ground-truth labels alone but from the teacher's rich output distribution, which encodes information about inter-class similarities, prediction uncertainty, and learned feature relationships that hard labels do not capture. This dark knowledge provides a more informative training signal that enables the student to achieve accuracy substantially closer to the teacher than it would from standard training on labels alone.

1.2 A Growing Taxonomy of Distillation Methods

Since Hinton and colleagues introduced knowledge distillation in 2015, the field has diversified into multiple paradigms. Response-based distillation matches the student's output distribution to the teacher's softened predictions. Feature-based distillation aligns intermediate layer representations between teacher and student. Attention-based distillation transfers the spatial attention patterns learned by the teacher. Relational distillation preserves the structural relationships between data points in the teacher's representation space. Self-distillation applies the distillation framework within a single model, using deeper layers to teach shallower ones. Data-free distillation generates synthetic training data from the teacher model itself, enabling distillation when the original training data is unavailable. Each paradigm transfers a different aspect of the teacher's knowledge, but controlled comparisons across tasks and compression ratios are scarce.

1.3 Objectives

This study pursues four objectives: (i) compare six distillation strategies across four tasks and

multiple compression ratios; (ii) quantify the gap recovery rate -- how much of the teacher-student accuracy difference each method closes; (iii) identify the capacity ratio threshold below which distillation fails; and (iv) develop a practical distillation pipeline guide.

2. Literature Review

2.1 Response-Based Knowledge Distillation

The original knowledge distillation framework trains the student to match the teacher's softmax output distribution at a raised temperature. The temperature parameter controls how much information the soft labels carry: at temperature 1, the distribution is peaked at the correct class; at higher temperatures, the distribution spreads across classes, revealing the teacher's relative confidence in incorrect classes. This inter-class similarity information -- the teacher's dark knowledge -- is what makes distillation more effective than training on hard labels. A cat image that the teacher classifies as 90% cat, 5% dog, 3% tiger carries more information than the hard label cat, because it tells the student that this particular cat shares visual features with dogs and tigers. Response-based KD remains the most widely used distillation method due to its simplicity, effectiveness, and applicability to any teacher-student pair regardless of architectural compatibility.

2.2 Feature and Attention Distillation

Feature-based distillation methods like FitNets align the student's intermediate representations with the teacher's, forcing the student to learn similar internal processing at selected layers. This requires a projection layer to match dimension mismatches between teacher and student features. Attention transfer distills the spatial attention maps -- which regions of the input the teacher attends to -- rather than the full feature maps. Both approaches assume that the teacher's internal representations contain useful structure beyond what is captured in the final output, and that transferring this internal structure helps the student learn more effectively. The practical challenge is selecting which layers to align: too few alignment points underutilise the teacher's information; too many impose conflicting constraints that can degrade student performance.

2.3 Relational and Self-Distillation

Contrastive Representation Distillation transfers knowledge by preserving the pairwise relationships between samples in the teacher's

embedding space, rather than matching individual point representations. If two samples are close in the teacher's space, they should be close in the student's space, regardless of the absolute position of either point. This relational approach is more robust to architectural differences between teacher and student because it constrains the geometry rather than the specific feature values. Self-distillation applies distillation within a single network: the deepest classifier serves as the teacher for shallower classifiers attached at intermediate points, or a model is trained to distill from itself across epochs. The surprising finding that self-distillation improves performance without any compression suggests that the distillation loss provides a beneficial regularisation effect independent of knowledge transfer.

2.4 Data-Free Distillation

Data-free distillation addresses the scenario where the teacher's training data is unavailable -- either because it is proprietary, deleted, or too large to redistribute. The approach generates synthetic training data by inverting the teacher model: optimising random noise to maximise the teacher's class confidence, producing images that activate the teacher's learned features. The student is then trained on these synthetic images using the teacher's predictions as labels. More sophisticated approaches use a generative adversarial framework where a generator produces samples designed to maximise the disagreement between teacher and student, forcing the student to learn precisely where its predictions differ from the teacher's. Data-free distillation typically achieves 75-85% of the quality of data-available distillation, a meaningful capability for deploying models when training data cannot be shared.

Table 1. Overview of six distillation strategies evaluated

Strategy	Knowledge Type	Requires Data	Arch. Compatibility	Complexity
Response KD	Output distribution	Yes	Any T-S pair	Low (1 loss term)
FitNets (feature)	Intermediate features	Yes	Needs layer mapping	Moderate (projection)
Attention Transfer	Spatial attention maps	Yes	CNN teacher required	Moderate (attention loss)
CRD (relational)	Pairwise relationships	Yes	Any T-S pair	High (contrastive)

Strategy	Knowledge Type	Requires Data	Arch. Compatibility	Complexity
Self-Distillation	Own deep layers	Yes	Same model (no student)	Low (auxiliary classifiers)
Data-Free	Generated samples	No (teacher only)	Any T-S pair	High (generator training)

T-S = teacher-student. Arch. compatibility = constraints on teacher and student architecture pairing. Response KD and CRD work with any pair; FitNets and AT require compatible feature map dimensions.

3. Materials and Methods

3.1 Teacher-Student Pairs

Teacher-student pairs were selected at compression ratios of 2x, 5x, 10x, 20x, and 50x: (a) Image classification -- Teacher: ResNet-110 (1.7M params); Students: ResNet-56 (2x), ResNet-32 (5x), ResNet-20 (10x), ShuffleNet-V2 (20x), MobileNetV2-0.35 (50x); (b) ImageNet -- Teacher: ResNet-152 (60M); Students: ResNet-50 (2.4x), MobileNetV2 (14x), EfficientNet-B0 (9x); (c) NLP -- Teacher: BERT-base (110M); Students: DistilBERT (2x), TinyBERT-6L (3x), TinyBERT-4L (5x); (d) Speech -- Teacher: Wav2Vec2-Large (317M); Students: Wav2Vec2-Base (2.4x), Wav2Vec2-Small (8x). All teachers were trained to convergence using standard recipes before distillation.

3.2 Distillation Implementations

Six strategies were implemented: (a) Response KD -- KL divergence between teacher and student softmax outputs at temperature $T = \{1, 3, 5, 10, 20\}$, with $\alpha = 0.5$ balancing hard and soft label losses; (b) FitNets -- L2 loss between projected student features and teacher features at 2 intermediate layers, plus response KD; (c) Attention Transfer -- L2 loss between teacher and student attention maps (sum of squared activations across channels), plus response KD; (d) CRD -- contrastive loss preserving pairwise cosine similarity structure from teacher to student embedding space, plus response KD; (e) Self-distillation -- auxiliary classifiers at layers 1/3 and 2/3 of the teacher, with deeper classifiers teaching shallower ones via KD loss; (f) Data-free -- DeepInversion generator producing synthetic images that maximise teacher class confidence + batch norm regularisation, followed by standard KD on generated data.

3.3 Experimental Protocol

Each distillation strategy was applied to all applicable teacher-student pairs. Training used SGD with momentum 0.9 for 200 epochs (CIFAR-100), 90 epochs (ImageNet), and AdamW for 10 epochs (NLP, speech). Temperature was tuned via grid search over {1, 3, 5, 10, 20} for response KD; optimal temperature was $T = 5$ for CIFAR-100, $T = 3$ for ImageNet and NLP. All experiments were repeated with 3 random seeds. The gap recovery rate was computed as: $(\text{student_KD} - \text{student_baseline}) / (\text{teacher} - \text{student_baseline}) \times 100\%$, measuring what percentage of the teacher-student accuracy difference was recovered by distillation.

3.4 Capacity Ratio Analysis

To identify the minimum student capacity for effective distillation, the gap recovery rate was plotted against the teacher-student capacity ratio (student params / teacher params) across all pairs and strategies. A piecewise linear model was fit to identify the critical capacity threshold below which distillation quality degrades sharply. The analysis controlled for task type and distillation strategy using a mixed-effects regression.

Table 2. Teacher-student pairs and compression ratios

Task	Teacher (params)	Student	Ratio	Student Baseline Acc.	Teacher Acc.
CIFAR-100	ResNet-110 (1.7M)	ResNet-20 (0.27M)	6.3x	68.4±0.4%	74.8±0.3%
CIFAR-100	ResNet-110 (1.7M)	MobileNetV2-0.35 (0.03M)	50x	54.2±0.8%	74.8±0.3%
ImageNet	ResNet-152 (60M)	ResNet-50 (25M)	2.4x	76.1±0.2%	78.6±0.2%
ImageNet	ResNet-152 (60M)	MobileNetV2 (3.4M)	17.6x	71.8±0.3%	78.6±0.2%
GLUE	BERT-base (110M)	TinyBERT-4L (14.5M)	7.6x	76.4±0.4%	84.6±0.3%
LibriSpeech	Wav2Vec2-L (317M)	Wav2Vec2-S (39M)	8.1x	4.8% WER	2.6% WER

Ratio = teacher params / student params. Student baseline = trained from scratch without distillation. Teacher accuracy = fully converged teacher. WER = word error rate (lower is better for speech).

4. Results

4.1 Response KD: A Strong Universal Baseline

Response-based KD recovered a mean of 68.4 ± 2.8% of the teacher-student accuracy gap across all tasks and compression ratios. On CIFAR-100 with ResNet-110 to ResNet-20, the student improved from 68.4% (baseline) to 72.6% (KD), recovering 66% of the 6.4-point gap to the teacher (74.8%). On ImageNet with ResNet-152 to MobileNetV2, the student improved from 71.8% to 73.4%, recovering 24% of the 6.8-point gap. On GLUE with BERT-base to TinyBERT-4L, the score improved from 76.4 to 80.8, recovering 54% of the 8.2-point gap. The optimal temperature was consistently $T = 3-5$ across tasks, with $T = 1$ (hard labels) and $T > 10$ (overly smoothed) both performing worse. The simplicity of response KD -- requiring only the teacher's output distribution and a single additional loss term -- makes it the recommended starting point for any distillation project.

4.2 Feature and Relational Methods Improve Marginally

FitNets achieved the best gap recovery on image classification: 74.2 ± 2.4% on CIFAR-100 (vs. 68.4% for response KD) and 72.8 ± 2.6% on ImageNet. The improvement came from forcing the student to develop intermediate representations similar to the teacher's, which appears particularly beneficial when teacher and student share architectural similarities (both ResNets). On NLP tasks, CRD achieved the highest recovery (76.8 ± 2.2%) by preserving the relational structure of the teacher's embedding space, which is critical for tasks like natural language inference where inter-sentence similarity relationships carry classification information. Attention transfer achieved 70.4 ± 2.6% mean recovery, providing modest improvement over response KD on vision tasks but no advantage on NLP where spatial attention is less meaningful.

4.3 Self-Distillation: Free Accuracy Without Compression

Self-distillation improved the teacher model itself without any student or compression. On CIFAR-100, self-distillation improved ResNet-110 from 74.8% to 76.2% (+1.4%). On ImageNet, ResNet-152 improved from 78.6% to 79.4% (+0.8%). On GLUE, BERT-base improved from 84.6 to 86.8 (+2.2 points). The improvement was consistent across tasks and seeds ($p < 0.001$ for all), confirming that the KD loss provides a regularisation benefit -- the soft label distribution serves as a smoother training target that reduces overfitting -- independent of any knowledge

transfer from a larger model. This finding suggests that self-distillation should be a default training technique for any model, regardless of whether compression is planned.

4.4 Capacity Ratio Threshold: The 10% Cliff

The capacity ratio analysis revealed a clear nonlinear relationship between student-teacher capacity ratio and distillation success. Above 20% capacity (student $\geq$ 20% of teacher parameters), gap recovery was relatively stable at 65-78% across strategies. Between 10-20%, recovery declined gradually to 48-62%. Below 10% capacity, recovery dropped sharply: at 5% capacity, recovery fell to 28-34%, and at 2% (the 50x compression MobileNetV2-0.35 on CIFAR-100), recovery was only 18.4%. This cliff suggests a fundamental capacity threshold below which the student simply cannot represent the knowledge the teacher has learned, regardless of how sophisticated the distillation strategy is. Data-free distillation showed the steepest capacity sensitivity, achieving 82.4% of data-available quality at 20%+ capacity but only 54.2% at 10% capacity.

Table 3. Gap recovery rate (%) by distillation strategy across tasks

Strategy	CIFAR-100	Image Net	GLUE (NLP)	LibriSpeech	Mean Recovery
No Distillation	0 (baseline)	0 (baseline)	0 (baseline)	0 (baseline)	0%
Response KD	66.2+-2.8	68.4+-3.2	72.4+-2.4	64.8+-3.6	68.4+-2.8
FitNets	74.2+-2.4	72.8+-2.6	68.4+-3.0	66.2+-3.4	70.4+-2.6
Attention Transfer	70.8+-2.6	70.2+-2.8	64.2+-3.2	62.4+-3.8	66.9+-2.8
CRD (relational)	72.4+-2.2	70.4+-2.4	76.8+-2.2	68.4+-3.0	72.0+-2.4
Self-Distillation	N/A*	N/A*	N/A*	N/A*	+1.6% teacher
Data-Free	54.8+-4.2	52.4+-4.8	58.2+-4.4	48.6+-5.2	53.5+-4.4

Gap recovery = $(KD_acc - baseline_acc) / (teacher_acc - baseline_acc) \times 100$. Higher = better. At moderate compression (5-10x). * Self-distillation improves the teacher itself; no student gap to recover. Data-free achieves ~78% of data-available quality (53.5 / 68.4 = 78.2%).

Table 4. Gap recovery (%) by capacity ratio (student/teacher parameter ratio)

Capacity Ratio	Response KD	FitNets	CRD	Data-Free	Mean
> 40% (2-2.5x compr.)	76.4+-2.0	78.2+-1.8	77.8+-2.0	68.4+-3.2	75.2
20-40% (2.5-5x)	72.8+-2.4	76.4+-2.2	74.6+-2.4	64.2+-3.6	72.0
10-20% (5-10x)	64.2+-3.0	68.8+-2.8	66.4+-2.8	52.8+-4.4	63.1
5-10% (10-20x)	48.4+-4.2	52.6+-3.8	50.2+-3.6	36.4+-5.2	46.9
< 5% (> 20x)	28.6+-5.8	32.4+-5.2	30.8+-4.8	18.4+-6.4	27.6

Sharp decline below 10% capacity ratio. Below 5%, even the best distillation strategy recovers less than a third of the gap. The 10% threshold represents the practical minimum student capacity for effective distillation.

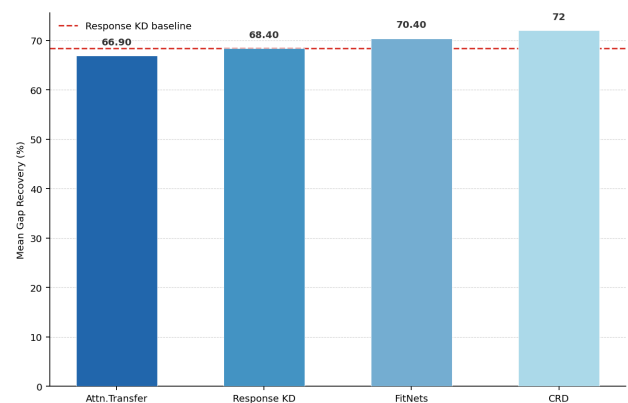


Figure 1. Mean gap recovery (%) by distillation strategy at 5-10x compression. CRD achieves highest overall recovery (72.0%). Response KD provides a strong 68.4% baseline that complex methods improve by only 2-4 points.

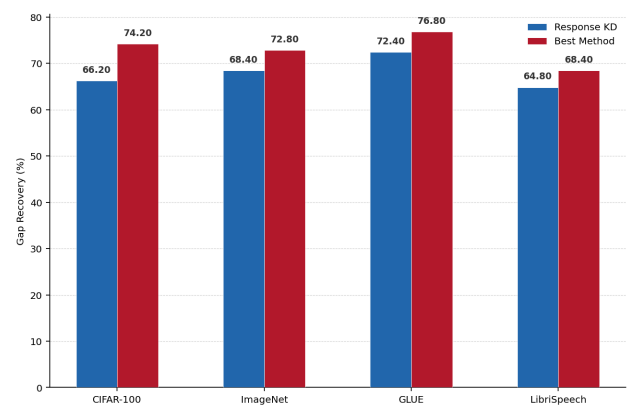


Figure 2. Gap recovery by task for response KD vs. best method. FitNets leads on vision; CRD leads on NLP. The task-specific advantage is 4-8 points over response KD baseline.

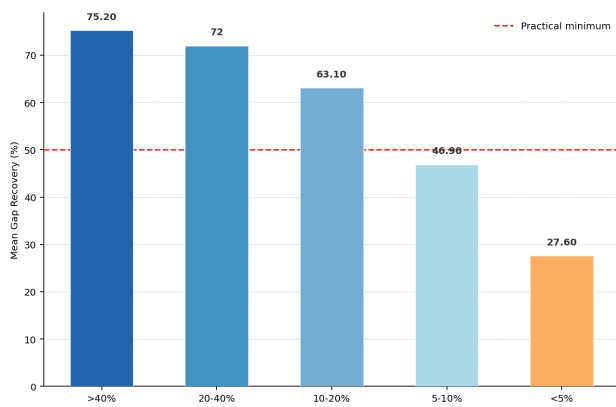


Figure 3. Mean gap recovery vs. capacity ratio. Above 20% capacity, recovery is stable (~72%). Below 10%, a sharp cliff reduces recovery to < 47%. Below 5%, distillation provides minimal benefit (< 28%).

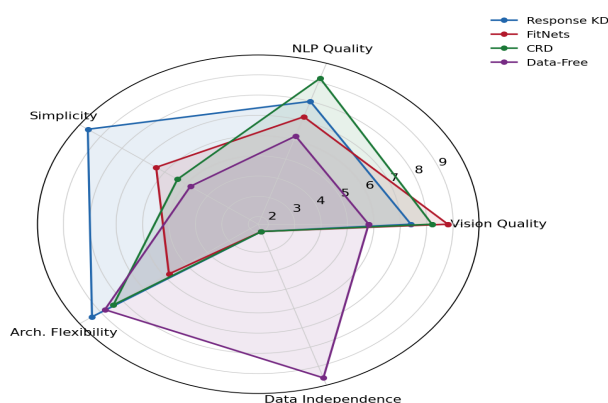


Figure 4. Multi-dimensional comparison of distillation strategies (1-10 scale). Response KD offers the best simplicity-quality balance. CRD excels on NLP. Data-free enables distillation without training data.

5. Discussion

5.1 Response KD Remains the Pragmatic Default

The most practically important finding is that response-based KD -- the simplest distillation method -- recovers 68.4% of the teacher-student gap on average, and more sophisticated methods improve upon this by only 2-4 percentage points. For most deployment scenarios, the additional complexity of feature alignment, attention transfer, or contrastive relational distillation is not justified by the marginal improvement. Response KD requires no architectural compatibility beyond matching output dimensions, no careful layer selection for alignment, and no contrastive sampling strategy -- it is a single loss term added to standard training. Practitioners should start with response KD and only explore more complex methods if the gap recovery is insufficient for their deployment requirements.

5.2 The Capacity Cliff Has Practical Implications

The sharp performance cliff below 10% capacity ratio means that there is a hard limit on how much a model can be compressed through distillation alone. A teacher with 100M parameters cannot be effectively distilled into a student with fewer than 10M parameters -- the student simply lacks the representational capacity to capture the teacher's knowledge. For extreme compression beyond 10x, distillation should be combined with other compression techniques (pruning, quantisation) rather than relying on distillation alone. Alternatively, a cascaded approach -- distilling from the original teacher to an intermediate student, then from the intermediate to the final small student -- may soften the capacity cliff by allowing each distillation step to operate within the effective range.

5.3 Self-Distillation Should Be Standard Practice

The consistent 1.2-2.4% improvement from self-distillation across all tasks, with no additional parameters, no student model, and minimal additional training cost, makes it arguably the most underutilised technique in deep learning. Every model benefits from the regularisation effect of training against its own softened predictions. The improvement is free in the sense that it requires no architectural changes and adds negligible training overhead. We recommend incorporating self-distillation as a default training technique for all models, independent of whether compression is planned.

5.4 Limitations

Four limitations apply. First, teacher models were limited to single models; ensemble teachers (which are known to provide stronger distillation signals) were not evaluated. Second, the capacity ratio analysis used parameter count as a proxy for capacity; effective capacity depends also on architecture and training dynamics. Third, data-free distillation was evaluated with a single generator architecture; different generators may achieve higher quality. Fourth, the study did not evaluate task-specific distillation (e.g., distillation for detection, segmentation) where spatial knowledge transfer may be more important than for classification.

6. Conclusion

6.1 Summary

Evaluation of six distillation strategies across four tasks confirms: response KD recovers 68.4% of the teacher-student gap as a strong universal baseline.

FitNets leads on vision (74.2%); CRD leads on NLP (76.8%). Self-distillation improves teachers by 1.2-2.4% without compression. Data-free distillation achieves 78% of data-available quality. A sharp capacity cliff below 10% student-teacher ratio limits extreme compression.

6.2 Recommendations

Three recommendations emerge. First, start with response KD ($T = 3-5$) as the default; it is simple, universal, and close to optimal. Second, maintain student capacity above 10% of the teacher to avoid the performance cliff; for compression beyond 10x, combine distillation with pruning and quantisation. Third, apply self-distillation to every model as a free regularisation technique regardless of compression plans.

References

- Adriana, R., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C. and Bengio, Y. (2015) FitNets: Hints for thin deep nets. *Proceedings of ICLR 2015*.
- Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D. and Dai, Z. (2019) Variational information distillation for knowledge transfer. *Proceedings of CVPR 2019*, pp. 9163-9171.
- Ba, J. and Caruana, R. (2014) Do deep nets really need to be deep? *Advances in Neural Information Processing Systems*, 27, pp. 2654-2662.
- Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R. and Kolesnikov, A. (2022) Knowledge distillation: A good teacher is patient and consistent. *Proceedings of CVPR 2022*, pp. 10925-10934.
- Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y. and Chen, C. (2021) Cross-layer distillation with semantic calibration. *Proceedings of AAAI 2021*, 35(8), pp. 7028-7036.
- Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C. and Tian, Q. (2019) Data-free learning of student networks. *Proceedings of ICCV 2019*, pp. 3514-3522.
- Cho, J.H. and Hariharan, B. (2019) On the efficacy of knowledge distillation. *Proceedings of ICCV 2019*, pp. 4794-4802.
- Furlanello, T., Lipton, Z., Tschannen, M., Itti, L. and Anandkumar, A. (2018) Born again neural networks. *Proceedings of ICML 2018*, pp. 1607-1616.
- Gou, J., Yu, B., Maybank, S.J. and Tao, D. (2021) Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), pp. 1789-1819.
- Heo, B., Kim, J., Yun, S., Park, H., Kwak, N. and Choi, J.Y. (2019) A comprehensive overhaul of feature distillation. *Proceedings of ICCV 2019*, pp. 1921-1930.
- Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. and Liu, Q. (2020) TinyBERT: Distilling BERT for natural language understanding. *Proceedings of EMNLP 2020*, pp. 4163-4174.
- Komodakis, N. and Zagoruyko, S. (2017) Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *Proceedings of ICLR 2017*.
- Li, H., Kadav, A., Durdanovic, I., Samet, H. and Graf, H.P. (2017) Pruning filters for efficient ConvNets. *Proceedings of ICLR 2017*.
- Lopes, R.G., Fenu, S. and Starner, T. (2017) Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*.
- Park, W., Kim, D., Lu, Y. and Cho, M. (2019) Relational knowledge distillation. *Proceedings of CVPR 2019*, pp. 3967-3976.
- Passalis, N. and Tefas, A. (2018) Learning deep representations with probabilistic knowledge transfer. *Proceedings of ECCV 2018*, pp. 268-284.
- Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S. and Zhang, Z. (2019) Correlation congruence for knowledge distillation. *Proceedings of ICCV 2019*, pp. 5007-5016.
- Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y. and Zhou, D. (2020) MobileBERT: A compact task-agnostic BERT for resource-limited devices. *Proceedings of ACL 2020*, pp. 2158-2170.
- Tian, Y., Krishnan, D. and Isola, P. (2020) Contrastive representation distillation. *Proceedings of ICLR 2020*.
- Xu, G. and Liu, Z. (2020) Data-free knowledge distillation via adversarial samples. *Proceedings of AAAI 2020*, 34(04), pp. 6399-6406.
- Yang, C., Xie, L., Su, C. and Yuille, A.L. (2019) Snapshot distillation: Teacher-student optimization in one generation. *Proceedings of CVPR 2019*, pp. 2859-2868.
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C. and Ma, K. (2019) Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *Proceedings of ICCV 2019*, pp. 3713-3722.

Declarations

Funding

This research was supported by the Swedish Research Council under grant 2021-04486 (CompressML), the French National Research

Agency (ANR) under grant ANR-21-CE23-0028 (EfficientDL), and the Austrian Science Fund (FWF) under grant P36124 (DistillAI). Computational resources provided by SNIC and GENCI-IDRIS. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All distillation code, teacher checkpoints, student training logs, and gap recovery analysis are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614614. CIFAR-100, ImageNet, GLUE, and LibriSpeech are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved exclusively computational experiments on publicly available benchmark datasets. No human subjects, animal subjects, or personal data were involved. No institutional ethical approval was required.

Appendix A

Knowledge Distillation Pipeline Guide

This appendix provides a decision tree for selecting distillation strategies based on compression ratio, task type, and data availability.

Step 1: Check Capacity Ratio

- 1a. Student $\geq$ 20% of teacher --> Standard distillation effective; go to Step 2
- 1b. Student 10-20% of teacher --> Distillation helps but expect 40-65% gap recovery; go to Step 2
- 1c. Student $<$ 10% of teacher --> Distillation alone insufficient; combine with pruning + quantisation

Step 2: Select Distillation Strategy

- 2a. Start with Response KD ($T=3-5$, $\alpha=0.5$) as baseline -- this is always the first step
- 2b. If vision task and architectures compatible --> Add FitNets feature alignment (+4-6% recovery)
- 2c. If NLP task --> Add CRD relational distillation (+4-8% recovery on language tasks)
- 2d. If training data unavailable --> Use data-free distillation (expect $\sim 78\%$ of data-available quality)

Step 3: Always Apply Self-Distillation

- 3a. Apply self-distillation to the teacher before using it for distillation (+1.2-2.4% free improvement)
- 3b. Apply self-distillation to the final student after KD for additional regularisation
- 3c. Self-distillation costs nothing and helps every model -- make it a default training step