Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(1) 2026

An Empirical Analysis of the Temperature Hyperparameter in Knowledge Distillation for Image Classification Student Networks

Hongbo Wang[1]，Yunqiang Sun[2,*]

[1] Computer Science, University of Southern California, CA, USA
[2]Computer Science and Technology, University of Jinan, Shandong, China

* Corresponding Email: sunyunqiang@qq.com

Abstract

Knowledge distillation transfers the softened class probabilities of a large teacher network to a compact student, and the temperature applied to those probabilities governs how much inter-class structure the student receives. While the temperature is widely treated as a routine hyperparameter, the magnitude and direction of its effect on a small student remain unevenly documented across the image-classification literature. This work reports a controlled empirical analysis that isolates the temperature while holding the teacher, the student, the loss weighting, and the optimisation budget fixed. A three-layer convolutional teacher (458,890 parameters) is distilled into a two-layer student (53,370 parameters) on Fashion-MNIST, and the temperature is swept across nine values from 1 to 20, each repeated over three random seeds. The distilled student attains its best top-1 accuracy of 85.20% at the lowest temperature, a moderate gain of 0.64 percentage points over the undistilled baseline of 84.56%, after which accuracy declines monotonically and falls below the baseline at the highest temperatures. Teacher-student prediction agreement follows the same descending pattern, while the entropy of the softened teacher targets rises steadily toward the uniform limit. The results indicate that, for a high-compression student under a fixed loss weight, restrained softening is preferable to aggressive softening, and they caution against treating large temperatures as a default choice.

Keywords: knowledge distillation, temperature hyperparameter, image classification, model compression

1. Introduction

1.1. Background and Motivation

Knowledge distillation has become a standard route to compressing image-classification networks, training a small student to reproduce the softened output of a larger teacher rather than

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

the one-hot labels alone [1]. The softening is controlled by a temperature that scales the logits before the softmax, raising the relative weight of non-target classes and exposing the similarity structure that the teacher has learned. A substantial body of work has since extended the idea across many forms of transferred knowledge, training schemes, and architectures [2]. The practical appeal is direct: compact convolutional classifiers suited to constrained deployment [3] often trail their larger counterparts, and distillation offers a way to narrow that gap without changing the deployed network.

The temperature sits at the centre of this mechanism, yet it is frequently fixed to a convention rather than examined. Reported defaults span a wide range, and the value that benefits one teacher-student pair need not benefit another. For a student with a large compression ratio, very soft targets may dilute the supervisory signal that the hard labels provide, while very sharp targets may withhold the inter-class information that motivates distillation in the first place. The direction of this trade-off, and its size relative to ordinary seed-to-seed variation, is what an experiment must establish.

Part of the difficulty is that the temperature is rarely studied in isolation. It is usually tuned jointly with the loss weighting, the teacher choice, and the optimisation schedule, so that a reported optimum reflects a particular combination rather than the temperature itself. When the temperature is reported at all, it tends to be a single chosen value rather than a curve, which leaves the shape of the accuracy response unobserved. A measurement that fixes everything else and varies only the temperature recovers that shape and shows whether the response is flat, single-peaked, or monotonic, each of which carries a different practical lesson for how much search effort the temperature deserves.

1.2. Scope and Contributions

This study restricts attention to a single question and answers it carefully rather than broadly. The teacher, the student, the loss weighting, and the training budget are all held constant, and only the temperature varies. The aim is a clean reading of how this one hyperparameter moves student accuracy, measured against an undistilled baseline and the teacher ceiling.

A. Research Questions

Three questions organise the analysis. The first asks whether distillation improves a high-compression student at all under a fixed loss weight, and by how much. The second asks how accuracy responds as the temperature increases across a representative grid. The third asks whether the agreement between student and teacher predictions tracks accuracy or diverges from it, which speaks to what the student is actually absorbing as the targets soften.

B. Summary of Contributions

The work contributes a reproducible single-variable temperature sweep with seed-level repetition, a joint reading of accuracy and teacher-student agreement against the entropy of the softened targets, and a cautious interpretation calibrated to the modest effect sizes that the measurements support. Every reported number derives from the runs described here, and the

spread across seeds is reported alongside each mean so that the reader can judge the reliability of the trend.

2. Related Work

2.1. Forms of Distilled Knowledge

A. Response-Based and Feature-Based Transfer

The original formulation transfers the teacher's output distribution, and later work broadened the channel of transfer to internal representations. Hint-based training aligns a student's intermediate features with those of a deeper teacher, allowing thin networks to benefit from mid-level guidance and to be trained in stages that first match features and then match outputs [4]. Attention transfer instead matches spatial attention maps derived from feature activations, passing on where the teacher concentrates its responses while leaving the exact activation values free [5]. These variants share the premise that the teacher holds information beyond the final logits, and they differ mainly in which signal they expose to the student. Because the feature signal is supplied in addition to the softened outputs rather than in place of them, the temperature continues to govern the output channel even when feature alignment carries part of the load.

B. Relational and Contrastive Objectives

A second line moves from individual outputs to the relations among examples. Relational distillation transfers distances and angles within the teacher's representation space rather than per-sample responses, encouraging the student to preserve the geometry of the embedding even where individual predictions differ [6]. Contrastive distillation frames the objective as preserving the structural agreement between teacher and student embeddings, drawing matched pairs together and pushing mismatched pairs apart [7]. A decoupled reformulation of the classical loss separates the target-class term from the non-target term and shows that the latter carries much of the useful signal, which clarifies why the softened distribution matters beyond its single most probable class [8]. Although these objectives reach beyond the temperature-scaled softmax, they continue to rely on softened comparisons, which keeps the temperature relevant to how strongly the non-target structure is expressed.

2.2. Training Schemes and Temperature-Related Studies

Distillation has also been recast as a training arrangement rather than a fixed teacher-to-student transfer. Mutual learning trains a cohort of peers that teach one another without a pre-trained teacher, with each network distilling from the running predictions of the others [9], while a teacher-assistant network bridges large capacity gaps by inserting an intermediate model so that no single transfer spans the full distance from teacher to student [10]. The temperature connects naturally to calibration, where a single scaling parameter adjusts the sharpness of predicted probabilities to match observed accuracy without altering the ranking of classes [11]. Related

analysis of label smoothing shows that softening the targets improves a network's own generalisation yet can erode the information a teacher passes to a student, because the smoothed teacher collapses the fine distinctions among non-target classes that distillation relies on [12]. That tension bears directly on the choice of distillation temperature, since raising the temperature softens the teacher's outputs in a comparable way. These studies frame temperature as consequential, while leaving its isolated effect on a compact student, traced across a full grid rather than at a single value, to be measured directly. The present study takes up that gap, treating the temperature as the object of measurement rather than a nuisance parameter to be tuned away.

3. Experimental Setup

3.1. Datasets and Networks

The experiments use Fashion-MNIST, a ten-class image-classification dataset whose categories include several visually similar garments, which makes the inter-class structure that distillation transfers meaningful rather than trivial. Pairs such as pullover, coat, and shirt overlap in appearance, so a well-trained teacher assigns graded probabilities across them rather than a single confident class, and that gradation is precisely the signal the temperature controls. A fixed subset of 12,000 training images is used so that the student operates in a moderate-data regime where the additional supervision from a teacher has room to help, and evaluation uses the full test set of 10,000 images so that every reported accuracy rests on the same large sample. The teacher is a three-block convolutional network with batch normalisation, of the kind used to approximate larger image classifiers [13], and the student is a shallower two-block convolutional network with no normalisation and narrower channels. The teacher holds 458,890 parameters and the student 53,370, a compression ratio of roughly 8.6 to 1, which places the student firmly in the regime where the supervisory signal matters and where a poor choice of temperature has somewhere to do harm. Holding both architectures fixed across the sweep means that any change in measured accuracy reflects the temperature rather than a change in capacity. Table 1 lists the two configurations.

Table 1. Teacher and Student Network Configurations

Network	Convolutional blocks	Channel widths	Parameters	Role
Teacher	3 (with batch normalisation)	32–64–64	458,890	fixed knowledge source
Student	2	16–16	53,370	distillation target

Data source: parameter counts measured directly from the network definitions used in this study.

3.2. Distillation Objective and Temperature

A. Softened Targets and the Loss

The teacher logits are converted to soft targets by a temperature-scaled softmax, in which dividing the logits by a temperature greater than one raises the entropy of the resulting distribution and increases the relative mass on non-target classes. The student is trained on a convex combination of two terms: a cross-entropy term against the ground-truth labels and a divergence term that aligns the student's temperature-scaled distribution with the teacher's. The soft term is multiplied by the square of the temperature so that its gradient magnitude stays comparable as the temperature changes, a correction that follows from the way the temperature enters the softmax derivative and that keeps the balance between the two terms stable across the sweep. Without this correction the soft term would shrink as the temperature rose, and the apparent effect of temperature would be confounded with a silent reduction in the weight of the distillation signal. The weight on the cross-entropy term is fixed at 0.3 throughout, so that any change in student behaviour can be attributed to the temperature alone rather than to a shifting loss balance. Fixing this weight is a deliberate restriction of scope: the interaction between temperature and loss weighting is real, yet folding it into the present sweep would prevent a clean reading of the temperature on its own.

B. Evaluation Metrics

Two quantities are recorded for each configuration. The first is top-1 test accuracy, the primary measure of student quality. The second is teacher-student fidelity, defined as the fraction of test images on which the student's top prediction matches the teacher's, which captures how closely the student reproduces the teacher's decisions independently of correctness. Reporting fidelity alongside accuracy separates two effects that are easily conflated, because a student can grow more faithful to the teacher without growing more accurate; the distinction between matching a teacher and generalising well has been emphasised in prior analysis of whether distillation achieves what it appears to [14]. The entropy of the teacher's softened targets is also computed at each temperature to characterise the supervisory signal that the student receives. Recording all three quantities at every temperature allows the accuracy trend to be read against its proposed cause rather than in isolation.

3.3. Protocol and Temperature Grid

A. Training Configuration

All networks are optimised with Adam at a learning rate of 0.001 and a batch size of 128. The teacher is trained once for eight epochs and then frozen, so that every student distils from an identical source. Each student is trained for five epochs, and the undistilled baseline uses the same architecture, optimiser, and budget while learning only from the ground-truth labels. The temperature grid contains nine values, 1, 2, 3, 4, 5, 6, 8, 12, and 20, chosen to sample the sharp, moderate, and strongly softened regimes. Every configuration, including the baseline, is repeated

across three random seeds that govern both initialisation and data ordering, and results are reported as the mean and standard deviation over those seeds. Table 2 collects the protocol.

Table 2. Training and Sweep Protocol

Setting	Value
Dataset	Fashion-MNIST (10 classes)
Training images	12,000
Test images	10,000
Optimiser	Adam, learning rate 0.001
Batch size	128
Teacher epochs	8
Student epochs	5
Cross-entropy weight	0.3
Temperature grid	1, 2, 3, 4, 5, 6, 8, 12, 20
Random seeds	3

Data source: configuration of the experiments conducted in this study.

B. Softening of Teacher Targets

Before turning to the student, it is useful to see how the temperature reshapes the signal it learns from. The mean entropy of the frozen teacher's predictive distribution over the test set climbs steadily as the temperature rises, from 0.178 nats at a temperature of 1 to 2.257 nats at a temperature of 20, the latter approaching the uniform-distribution ceiling of ln 10 = 2.303 nats. At the highest temperature the teacher's outputs are very close to uniform, leaving little discriminative structure for the student to imitate, while at the lowest temperature the distribution stays sharp and the soft term behaves much like a confident relabelling that mostly reinforces the true class. The increase is steepest between temperatures of 1 and 4, where each unit of temperature adds substantial entropy, and it flattens thereafter as the distribution nears uniformity and further softening changes little. This concave progression matters for interpreting the student results, because it means the early part of the grid is where the supervisory signal changes most, and it sets the expectation that any temperature effect on the student should be most pronounced in that same low range. Figure 1 traces this progression and frames the student results that follow.

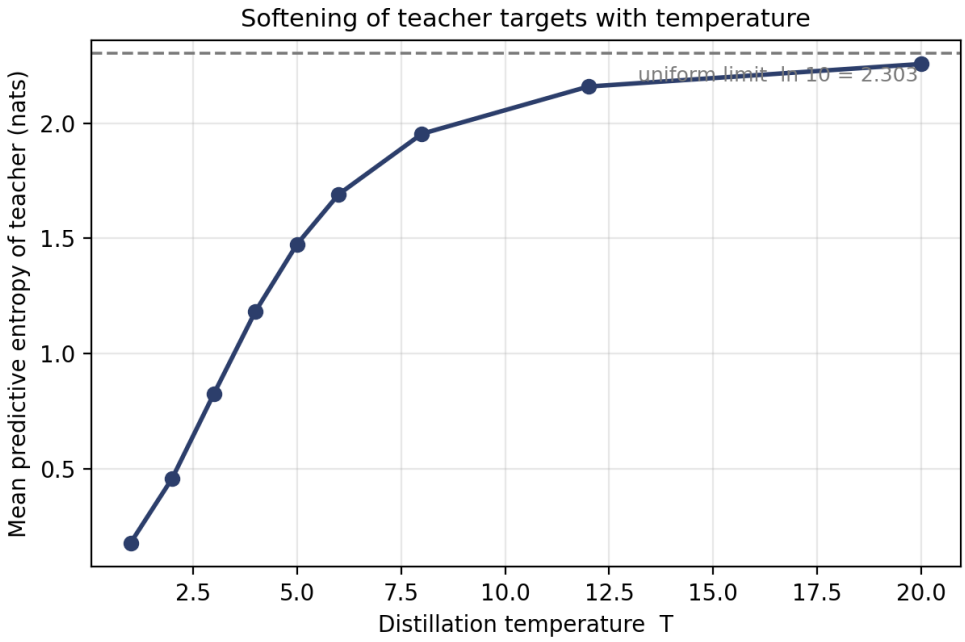


Figure 1. Entropy of the Softened Teacher Targets Across Temperatures

Mean predictive entropy of the frozen teacher over the 10,000-image test set as a function of the distillation temperature. The entropy rises monotonically from 0.178 nats at a temperature of 1 to 2.257 nats at a temperature of 20, approaching the uniform-distribution limit of ln 10 = 2.303 nats marked by the reference line. The steep early rise between temperatures of 1 and 4 shows that most of the softening occurs in the low-temperature range, after which the distribution flattens toward uniformity and the discriminative structure available to the student diminishes.

4. Results and Analysis

4.1. Accuracy Across Temperatures

A. Distillation Versus the Baseline

The undistilled student reaches a mean top-1 accuracy of 84.56% with a standard deviation of 0.47 across seeds, and the teacher reaches 89.68%, so the capacity gap leaves clear room for transfer. Distillation closes part of this gap, but only at restrained temperatures. At a temperature of 1 the distilled student attains 85.20%, a moderate improvement of 0.64 percentage points over the baseline, and at a temperature of 2 it attains 84.74%, an improvement of 0.18 points. The gain is real yet small relative to the teacher's advantage, which is consistent with the difficulty of transferring across a large compression ratio noted in earlier capacity studies [15]. Table 3 reports the full set of accuracies with their deviations and the difference from the baseline.

Table 3. Top-1 Test Accuracy Across Temperatures (mean ± standard deviation over three seeds)

Configuration	Top-1 accuracy (%)	Δ vs baseline (pp)
Baseline (no distillation)	84.56 ± 0.47	—
T = 1	85.20 ± 0.44	+0.64
T = 2	84.74 ± 0.52	+0.18
T = 3	84.55 ± 0.58	−0.01
T = 4	84.51 ± 0.63	−0.05
T = 5	84.50 ± 0.45	−0.06
T = 6	84.26 ± 0.62	−0.30
T = 8	84.06 ± 0.55	−0.50
T = 12	84.07 ± 0.36	−0.49
T = 20	83.93 ± 0.31	−0.63
Teacher (reference)	89.68	+5.12

Data source: test-set evaluation of the distilled students trained in this study.

B. The Temperature–Accuracy Trend

The pattern across the grid is orderly. Accuracy is highest at a temperature of 1, drifts down through the moderate range where the distilled student is statistically indistinguishable from the baseline near temperatures of 3 to 5, and continues to fall at high temperatures until it sits 0.63 points below the baseline at a temperature of 20. The monotonic descent, visible in Figure 2, indicates that for this high-compression student the benefit of distillation is concentrated at sharp targets and is gradually surrendered as softening increases. The finding stands in contrast to methods that deliberately raise or learn the temperature during training [16], and it suggests that such gains depend on factors beyond the static temperature studied here. The standard-deviation band overlaps the baseline through much of the moderate range, so the practical reading is that aggressive softening offers no advantage and a measurable penalty at the extreme. The intermediate values make the point concretely: at temperatures of 3, 4, and 5 the distilled student lands within six hundredths of a percentage point of the baseline, so distillation there neither helps nor harms, and only the two sharpest settings clear the baseline by a margin worth claiming. The descent is gradual rather than abrupt, with no single temperature at which performance collapses, which is consistent with a steady loss of structure as the targets soften rather than a threshold effect.

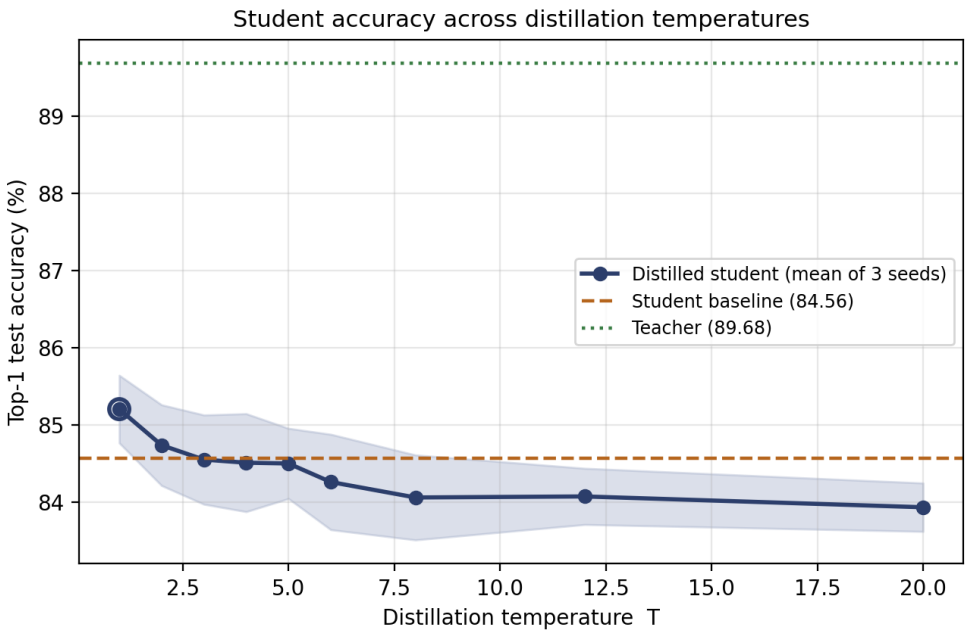


Figure 2. Student Accuracy as a Function of Distillation Temperature

Mean top-1 test accuracy of the distilled student across nine temperatures, with the shaded band spanning one standard deviation over three seeds. The dashed reference marks the undistilled baseline at 84.56% and the dotted reference marks the teacher at 89.68%. Accuracy peaks at 85.20% at a temperature of 1, declines through the moderate range, and reaches 83.93% at a temperature of 20, dropping below the baseline beyond a temperature of about 5. The circled point at a temperature of 1 denotes the best configuration, and the gap to the teacher line shows how much of the capacity difference distillation leaves unrecovered.

4.2. Fidelity and Target Softening

Teacher-student fidelity, the rate at which the student's top prediction matches the teacher's, behaves much like accuracy and declines as the targets soften. Fidelity is highest at 89.21% at a temperature of 1, eases to 88.50% at a temperature of 4, and settles near 88.2% to 88.3% at the highest temperatures, as Table 4 records together with the teacher target entropy. The joint movement is informative: as the entropy of the teacher's targets rises and the discriminative structure thins, the student both agrees with the teacher less often and classifies less accurately. Reading fidelity and accuracy side by side shows that softening does not buy closer imitation at the cost of accuracy; it reduces both at once. The decline in fidelity is gentler than the decline in accuracy, easing by roughly one percentage point across the full grid while accuracy falls by about 1.3 points from its peak, which indicates that the student keeps tracking the teacher's coarse decisions even as its own correctness slips. The mild levelling of fidelity at the highest temperatures, where it settles near 88.3% rather than continuing to fall, mirrors the flattening of the teacher entropy and suggests that once the targets are nearly uniform, further softening changes little for either quantity. This coupling is consistent with the view that an exact match of teacher

and student logit scales is neither necessary nor uniformly helpful, a concern raised in work that standardises logits before applying the temperature [17]. Figure 3 presents the fidelity trend and its relationship to accuracy across the grid.

Table 4. Teacher Target Entropy and Teacher–Student Fidelity Across Temperatures

Temperature T	Teacher target entropy (nats)	Teacher–student fidelity (%)
1	0.178	89.21
2	0.459	88.74
3	0.825	88.46
4	1.183	88.50
5	1.475	88.70
6	1.691	88.39
8	1.954	88.20
12	2.159	88.28
20	2.257	88.29

Data source: entropy of the frozen teacher and top-1 agreement of the distilled students measured in this study; the uniform-distribution entropy is ln 10 = 2.303 nats.

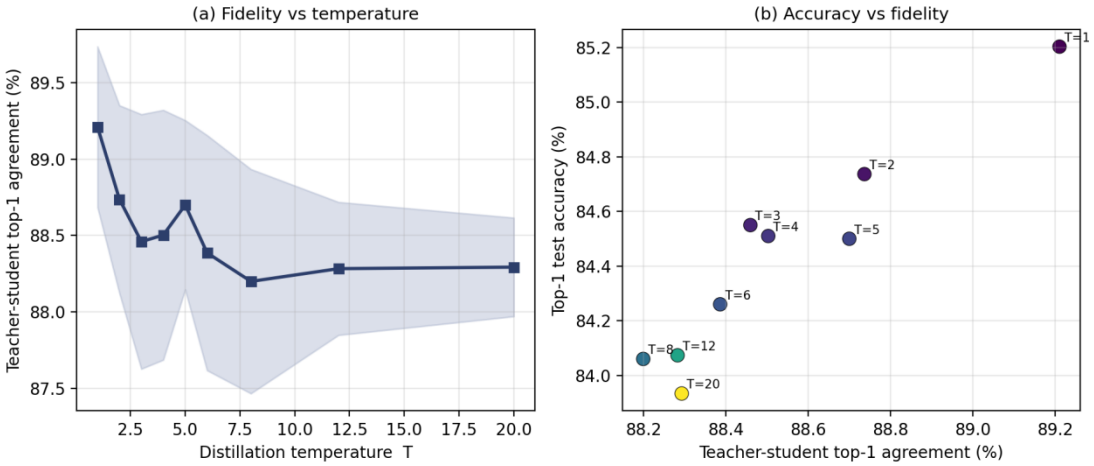


Figure 3. Teacher–Student Fidelity and Its Relationship to Accuracy

Distillation fidelity across temperatures and its link to student accuracy. (a) Mean top-1 agreement between student and teacher declines from 89.21% at a temperature of 1 to about 88.2% at a temperature of 8, with the shaded band spanning one standard deviation over three seeds. (b) Student accuracy plotted against fidelity, with each point labelled by its temperature, shows that the most faithful configurations are also the most accurate, and that both quantities deteriorate together as the temperature increases. The shared downward movement indicates that softening past the sharp regime removes useful structure rather than trading accuracy for imitation.

4.3. Interpretation

A. Why Lower Temperatures Help Here

The evidence points to a single coherent account. At low temperatures the teacher's targets remain sharp and carry a confident, largely correct relabelling that reinforces the ground-truth signal, while still exposing the relations among the nearest competing classes. As the temperature grows, the targets flatten toward uniformity, the entropy approaches its ceiling, and the soft term increasingly asks the student to match a near-uniform distribution that conveys little about class identity. Under a fixed cross-entropy weight the student cannot compensate for this dilution, and the accuracy decline follows the entropy rise closely, with the steepest losses appearing once the teacher entropy passes the midpoint of its range. For a student compressed by nearly an order of magnitude, the supervisory value of the teacher lies in restrained rather than aggressive softening. Schedules that move the temperature during training, annealing from softer to sharper targets as the student matures, implicitly acknowledge this regime dependence by spending more of the training budget where the signal is most informative [18], and the static measurements here help explain why the sharp end of the range is the productive one to favour.

B. Limits of the Observed Effect

The size of the effect deserves the same emphasis as its direction. The best improvement over the baseline is 0.64 percentage points, the standard-deviation bands overlap across neighbouring temperatures, and the moderate range is statistically close to the baseline. The safe reading is that temperature choice protects against a downside at high values rather than unlocking a large upside at low ones, at least for this teacher, this student, and this fixed loss weight. A larger teacher, a different compression ratio, or a different cross-entropy weight could shift the location of the peak, and the present measurements bound the claim to the regime actually tested rather than extending it to settings that were not. The honest reading of the effect size is that the temperature here acts as a guardrail more than a lever: setting it well prevents a loss of up to 0.63 points, while the best it adds over the baseline is 0.64 points, so the downside avoided and the upside gained are of similar small magnitude.

5. Discussion

5.1. Practical Implications

For a practitioner distilling into a small image classifier under a fixed loss balance, the measurements argue for caution with the temperature. A sharp setting near a temperature of 1 to 2 captured the available benefit, while temperatures beyond about 5 returned students that were no better than, and eventually worse than, training without a teacher at all. Treating a large temperature as a safe default would have cost accuracy in this setting, and the cost grew steadily with the temperature rather than appearing only at the extreme, which means the penalty is hard to notice from a single high-temperature run that lacks a baseline for comparison. Where compute

permits only a narrow search, concentrating that search at the sharp end of the range is the more productive strategy, since the upper half of the grid offered nothing the lower half did not. Reporting the seed-level spread is essential because the differences between adjacent temperatures are comparable to the variation between seeds, and a single run can easily suggest a peak that does not survive repetition. The practical takeaway is modest by design: the temperature is worth setting deliberately, but the gain it offers a high-compression student is small and is easily erased by softening too far. A reasonable default for this kind of student is to begin at a sharp temperature and to widen the search only if a baseline comparison reveals headroom that the sharp setting leaves unclaimed.

5.2. Limitations and Future Work

The study is deliberately narrow, and its limits define the natural next steps. A single dataset, one teacher-student pair, a compact training budget, and a fixed cross-entropy weight were chosen to isolate the temperature, and each of these choices constrains how far the conclusion travels. A dataset with many more classes would enlarge the space of non-target relations that softening exposes, and might move the peak away from the sharpest setting if the additional structure rewards a gentler distribution. A smaller compression ratio would give the student more capacity to absorb soft targets, while a larger and more accurate teacher might supply cleaner structure that tolerates more softening. Extending the sweep to additional datasets, to teacher-student pairs of varying compression ratios, and to a joint exploration of temperature and loss weight would test whether the monotonic decline observed here is specific to a high-compression student or a more general property. Longer training schedules would clarify whether the modest gain widens with more optimisation, and repeating the protocol with more seeds would tighten the bands enough to resolve the small differences that the present spread leaves uncertain. These extensions would build on the controlled single-variable design used here while broadening the conditions under which the temperature's role is understood, and each would keep the same discipline of changing one factor at a time so that the resulting curves remain interpretable. A further direction is to record not only final accuracy but the trajectory of fidelity through training, which would reveal whether the temperature shapes how quickly the student aligns with the teacher as well as where that alignment settles.

References

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.*

[2] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*(6), 1789–1819.

[3] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520).

[4] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for

thin deep nets. In *International Conference on Learning Representations.*

[5] Zagoruyko, S., & Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations.*

[6] Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3967–3976).

[7] Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive representation distillation. In *International Conference on Learning Representations.*

[8] Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11953–11962).

[9] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4320–4328).

[10] Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., & Ghasemzadeh, H. (2020). Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 5191–5198.

[11] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330).

[12] Müller, R., Kornblith, S., & Hinton, G. (2019). When does label smoothing help? In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 4694–4703).

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

[14] Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., & Wilson, A. G. (2021). Does knowledge distillation really work? In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 6906–6919).

[15] Cho, J. H., & Hariharan, B. (2019). On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4794–4802).

[16] Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J., & Yang, J. (2023). Curriculum temperature for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(2), 1504–1512.

[17] Sun, S., Ren, W., Li, J., Wang, R., & Cao, X. (2024). Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15731–15740).

[18] Jafari, A., Rezagholizadeh, M., Sharma, P., & Ghodsi, A. (2021). Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2493–2504).