Revisiting Label Smoothing and Knowledge Distillation Compatibility: What was Missing?

Keshigeyan Chandrasegaran, Ngoc-Trung Tran *, Yunqing Zhao *, Ngai-Man Cheung

{ keshigeyan, ngaiman_cheung }@sutd.edu.sg

Singapore University of Technology and Design (SUTD)

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

ICML International Conference On Machine Learning

Introduction

Background: This work investigates the compatibility between Label Smoothing (LS) and Knowledge Distillation (KD). Contemporary works studying this thesis statement take contradictory standpoints.

Does LS in a teacher network suppress the effectiveness of KD?
Müller et al. (2019) : ● "If a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective." ● "Label smoothing can hurt distillation."

Shen et al. (2021) : ● "Label smoothing will not impair the predictive performance of students."
● "Label smoothing is compatible with knowledge distillation."

Our Contributions: Our contributions are the discovery, analysis and validation of *systematic diffusion* as the missing concept which is instrumental in understanding / resolving these contradictory findings.

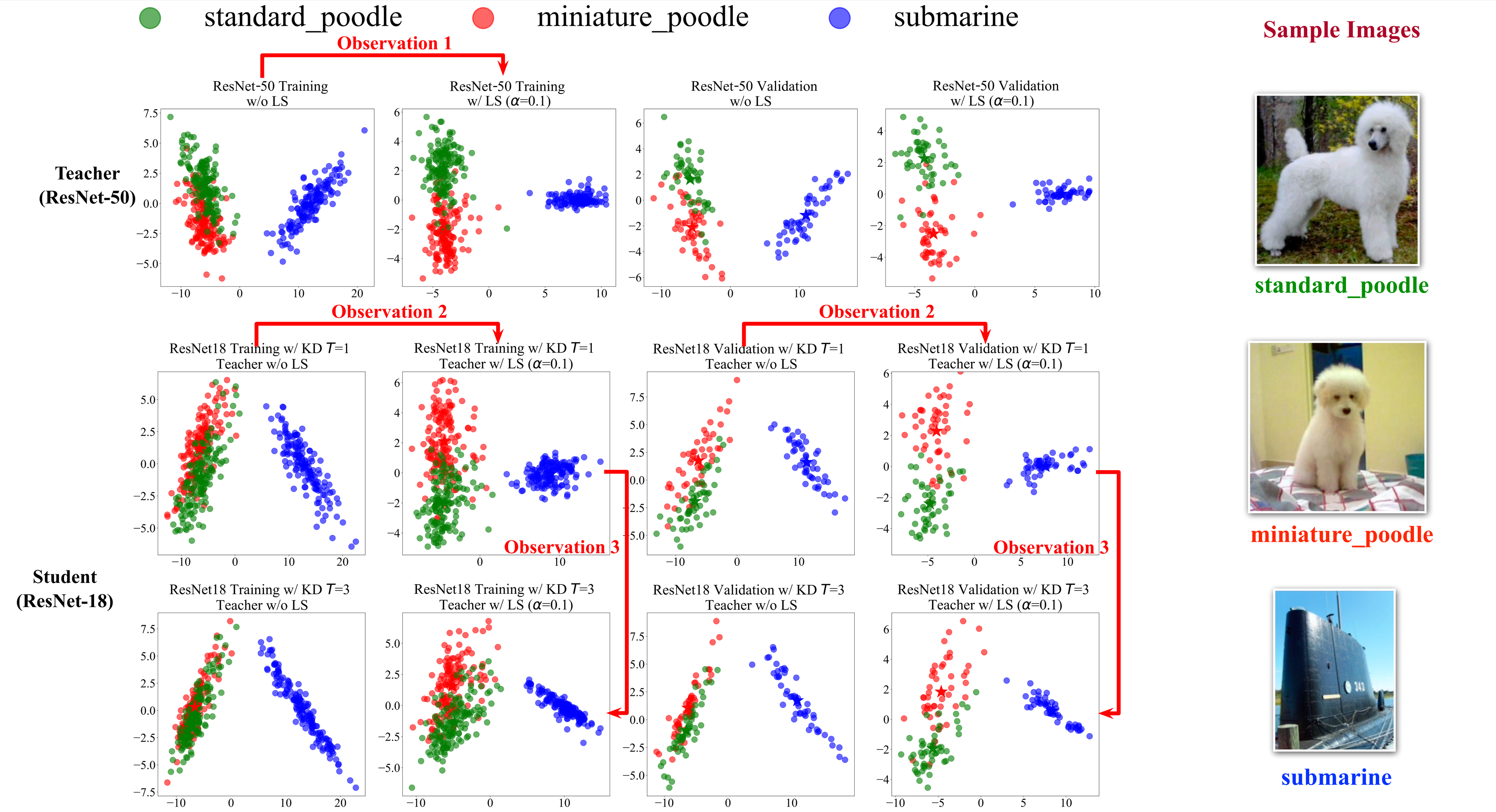
Systematic Diffusion in Student : In the presence of an LS-trained teacher, KD at higher temperatures systematically diffuses penultimate layer representations learnt by the student towards semantically similar classes. This *systematic diffusion* essentially curtails the *distance enlargement benefits* of distilling from a LS-trained teacher, thereby rendering KD at increased temperatures ineffective.

We show this *systematic diffusion* qualitatively by visualizing penultimate layer representations, and quantitatively using our proposed relative distance metric called diffusion index (η).

Our Main Findings on LS and KD Compatibility

		Information Erasure (Incompatibility)	Distance enlargement (compatibility)	**Systematic Diffusion (Incompatibility)**	**Conclusion**
Müller et al. 2019		LS erases relative information in the logits			LS-trained teacher can hurt KD
Shen et al. 2021		With LS, some relative information in the logits is still retained	LS enlarges the distance between semantically similar classes		Benefits outweigh disadvantages. LS is compatible with KD.
Our work	Lower T (i.e.: $T = 1$)	We agree with Shen et al., 2021 in information erasure	We validate the inheritance of distance enlargement in the student (Not shown in prior works)	With KD of lower T (i.e.: T=1), there is lower degree of systematic diffusion. This doesn't curtail the distance enlargement benefit.	At lower levels of systematic diffusion in student, LS is compatible with KD
	Increase of T	The loss of logits relative information cannot be recovered with an increased T	We agree with Shen et al., 2021 observation, but the distance enlargement is curtailed at an increased T.	With KD of increased T, there is systematic diffusion of penultimate representations towards semantically similar classes, curtailing the distance enlargement benefits.	At higher levels of systematic diffusion in student, LS and KD are not compatible.

Systematic Diffusion in Student



Visualization of the penultimate layer representations (Teacher=ResNet-50, Student=ResNet-18, Dataset=ImageNet-1K). We follow previous works and use three-class analysis: two semantically similar classes (**standard_poodle**, **miniature_poodle**) and one semantically dissimilar class (**submarine**).

Main Observations

Observation 1: The use of LS on the teacher leads to tighter clusters and erasure of logits' information as claimed by Müller et. Al (2019). In addition, increase in central distance between semantically similar classes (**standard_poodle**, **miniature poodle**) as claimed by Shen et al. (2021) can be observed.

Observation 2: We further visualize the student's representations. Increase in central distance between semantically similar classes can also be observed. This confirms the transfer of this benefit from the teacher to the student.

Observation 3 (Our main discovery): KD at an increased T causes *systematic diffusion* of representations between semantically similar classes (**standard_poodle**, **miniature_poodle**). This curtails the central distance enlargement benefits between semantically similar classes due to the use of an LS-trained teacher.

Quantifying Systematic Diffusion: η measurements

The principal idea of this metric is to quantify the distance change between clusters in the student when distilled from an LS-trained teacher at higher T.

See Section 4 in Paper for more details

Set 1 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
Chesapeake Bay retriever	-0.392	0.162	-1.082	0.269
curly-coated retriever	-0.578	0.179	-2.024	0.383
flat-coated retriever	-1.729	0.380	-3.320	0.655
golden retriever	-0.880	0.228	-2.594	0.555
Labrador retriever	-2.758	0.501	-4.618	0.840

Set 2 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
thunder snake	-2.316	0.376	-3.584	0.511
ringneck snake	-0.463	0.058	-0.757	0.094
hognose snake	-1.528	0.258	-4.067	0.631
water snake	-2.028	0.326	-3.053	0.478
king snake	-2.474	0.521	-4.577	0.840

Set 1 : ResNet-50 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
Chesapeake_Bay_retriever	-1.061	0.180	-1.346	0.240
curly-coated_retriever	-0.764	0.127	-1.193	0.207
flat-coated_retriever	-0.983	0.169	-0.331	0.056
golden_retriever	-0.744	0.159	-0.911	0.182
Labrado_retriever	-1.336	0.236	-1.468	0.257

Set 2 : ResNet-50 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
thunder snake	-2.565	0.417	-0.778	0.105
ringneck snake	-2.224	0.358	-0.726	0.102
hognose snake	-3.748	0.623	-2.173	0.342
water snake	-1.631	0.258	-0.390	0.037
king snake [2]	-1.969	0.339	0.956	-0.159

Main Experiments

Standard Image Classification (ImageNet-1K)
ResNet-50 to ResNet-18, ResNet-50 KD

	T \ α	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	76.130 / 92.862	76.196 / 93.078
Student : ResNet-18	$T = 1$	71.547 / 90.297	**71.616 / 90.233**
	$T = 2$	71.349 / 90.359	68.428 / 89.139
	$T = 3$	69.570 / 89.657	66.570 / 88.631
	$T = 64$	66.230 / 88.730	65.472 / 89.564
Student : ResNet-50	$T = 1$	76.502 / 93.059	**77.035 / 93.327**
	$T = 2$	76.198 / 92.987	76.101 / 93.115
	$T = 3$	75.388 / 92.676	**75.821 / 93.065**
	$T = 64$	74.291 / 92.399	**74.627 / 92.639**

Fine-grained Image Classification (CUB200-2011)
ResNet-50 to ResNet-18, ResNet-50 KD

	T \ α	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	81.584 / 95.927	82.068 / 96.168
Student : ResNet-18	$T = 1$	80.169 / 95.392	**80.946 / 95.312**
	$T = 2$	80.808 / 95.593	80.428 / 95.518
	$T = 3$	80.785 / 95.674	78.196 / 95.213
	$T = 64$	73.611 / 94.529	67.161 / 93.062
Student : ResNet-50	$T = 1$	82.902 / 96.358	**83.742 / 96.778**
	$T = 2$	82.534 / 96.427	**83.379 / 96.537**
	$T = 3$	82.091 / 96.243	**82.142 / 96.427**
	$T = 64$	79.784 / 95.927	77.206 / 95.812

We show top1 / top5 accuracies for Image classification (standard, fine-grained) KD experiments

Extended Experiments

Compact Student Distillation (CUB200-2011)
ResNet-50 to MobileNet-V2

	T \ α	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	81.584 / 95.927	82.068 / 96.168
Student : MobileNet-V2	$T = 1$	81.144 / 95.677	**81.731 / 95.754**
	$T = 2$	81.895 / 95.858	80.609 / 95.47
	$T = 3$	81.257 / 95.677	78.961 / 95.306
	$T = 64$	75.441 / 94.702	70.435 / 93.494

We show top1 / top5 accuracies for KD experiments

Neural Machine Translation (IWSLT, English -> German)
Transformer to Transformer

	T \ α	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : Transformer	-	26.461	26.750
Student : Transformer	$T = 1$	24.914	**25.085**
	$T = 2$	23.103	**23.421**
	$T = 3$	21.999	**22.076**
	$T = 64$	6.564	6.461

We show BLEU scores for KD experiments

Key Takeaways

Systematic Diffusion in Student: In the presence of an LS-trained teacher, KD at higher temperatures *systematically* diffuses penultimate layer representations learnt by the student towards semantically similar classes. This *systematic diffusion* essentially curtails the benefits of distilling from an LS-trained teacher, thereby rendering KD at increased temperatures ineffective.

Our discovery on *systematic diffusion* was the missing concept that is instrumental in resolving the contradictory findings of Müller et al. 2019 and Shen et al. 2021, thereby establishing a foundational understanding on the compatibility between LS and KD.

A rule of thumb for practitioners: We suggest using an LS-trained teacher with a low-temperature transfer (i.e., $T = 1$) to render high performance students.

* Equal Contribution

References

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems, 32*.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

Code & Pre-trained Models