Efficiently Robustify Pre-Trained Models
—Supplementary Material—

Nishant Jain [*]
IIT Roorkee

Harkirat Behl
Microsoft Research

Yogesh Singh Rawat
CRCV, UCF

Vibhav Vineet
Microsoft Research

Abstract

This supplementary contains more experiments involving ablations or analysis on various datasets, to further provide better understanding of our method as well as its advantages. We provide analysis of various ablations in our method in Sec. 1. Following this we provide further analysis of knowledge transfer for a given student w.r.t. various teachers in Sec. 2. We also show more results of our approach against the baselines, under transfer learning setup, in Sec. 3 along with extending it to other multi-modalities as well in Sec. 4. Finally, we further analyze our scheme and baselines on ImageNet-P dataset in Sec. 5 followed by amount of parameters we tune for different students (Sec. 6) and implementation details (Sec. 7) for reproducibility.

1. Further Ablations

We further provide an analysis of various variants, so as to understand importance of each of the proposed modules namely the *multi-headed architecture*, *knowledge distillation* from small to big network and *uncertainty/kl divergence* used at the inference time. We begin with analyzing the importance of proposed inference procedure.

1.1. Inference

We first begin with ablation on the proposed scheme for inference involving *Monte-Carlo Dropout* (MCD) uncertainty $\mathcal{U}_{mc}$ and KL divergence calculation (Sec. 4.3.2 in the paper). We also define a new term along with accuracy, for analyzing these inference time ablations. It is the fraction of examples in the test set, for a given dataset , which are assigned the correct head (clean for in-distribution and unclean for distribution/dataset shift). We denote it by $F_{correct}$. We start by analyzing the importance of KL divergence and thereby comparing our proposed scheme against the variant without any KL divergence calculation at the inference time.

KL Divergence Table 1 shows the analysis of our method with and without (*w/o*) KL divergence calculation at the inference time, on all the distribution shift datasets used in the paper for Visual Evaluation setting. The variant without the KL divergence term at the inference time is denoted as *Ours w/o KLD* in the table. It also shows the results for transfer learning experiments. The student model corresponds to multi-modal CLIP *ViT-L@333px* network and a single modal ViT-B/16 is used as the teacher. The last column of the table consists of average accuracy across all the distribution shift datasets. It can be observed that there is a small gain (average of 0.9%) across all the distribution shift scenarios when using our complete method as compared to this variant. Thus, using KL divergence at the inference time further rectifies the model performance. Also, Table 2 shows the analysis of $F_{correct}$ metric on all the distribution shift datasets used. Here also, a noticeable difference is observed implying KL divergence is significantly helping in deciding the correct head for the input. Next, we consider the ablation for the $\mathcal{U}_{mc}$ term.

Uncertainty. Similar analysis for $\mathcal{U}_{mc}$ is shown in Table 1 and the row corresponding to *Ours w/o* $\mathcal{U}_{mc}$ denotes this case. Again, removing the uncertainty component causes a depreciation in performance, but more than removing the KL divergence component (average decrease of 1.9% as compared to 0.9% from removing KL Divergence term at the inference time). Hence, both components are necessary for most optimal prediction. Furthermore, from Table 2 showing analysis of $F_{correct}$, it can be observed that using this $\mathcal{U}_{mc}$ term is significantly helping in deciding the correct head, again impacting more than the KL Divergence term.

Confidence based head selection. We also compare our head selection scheme with a case where instead of KL divergence or uncertainty, the predictive confidence of the clean and unclean heads is used for selecting the final classification head. Table X also contains the results for this method in the row *Ours(max logit)*. Again our proposed head selection scheme surpasses it with a significant margin.

[*] Correspondence to Nishant Jain at njain@cs.iitr.ac.in.

1.2. Knowledge Distillation

We now discuss the importance of the knowledge distillation (KD) module (Sec. 4.3 in the paper) proposed in our work, used for tuning the student model parameters. For this, we define another variant of our method with the multi-headed architectural scheme but without any KD. Table 1 also shows the analysis for this case using same models and scenarios as the above cases, compared against our method. The row corresponding to *Ours w/o K.D.* corresponds to this ablation. It can be observed that our method improves performance significantly as compared to this ablation (average 2.5% across distribution shifts), thereby showing utility of knowledge transfer. On clean dataset, only a marginal increase in performance is observed. Thus, the knowledge distillation component in our method plays a significant role in inducing robustness. Further comparison against this ablation for different teacher student pairs is shown in table 3. Again significant performance improvement, upto 2.6%, can be observed for our method using knowledge distillation.

1.3. Amount of dataset

We further analyze the impact of using different amounts of data while distilling knowledge, as it is crucial in deciding the computational overhead. Let us denote the total number of examples in the augmented data as $N^{'}$ and the number of examples (sampled randomly) for updating the student model as $aN^{'}(0 < a < 1)$. Here, we analyze our scheme for different values of a when it is applied on the ResNet-101 student with ResNet-34 as the teacher. The results are shown in Figure 1. It can be observed that as the data for tuning increases, the performance gap between our method and baseline increases, implying robustness of our scheme increases significantly w.r.t. data as compared to baseline.

2. Analyzing improvements in Teachers v/s Students

We now analyze the effect of improving robust accuracy of the small teacher model on the robustness it transfers to the student model using our method. For this, we analyze the ImageNet-C accuracy of the robustified student when using different teachers (having different ImageNet-C accuracies after their robustification). We fix the student to be a CLIP ViT-B/16 model and use robustified RN-34, RN-50, RN-101, RN-151, ViT-Tiny, ViT-Small as the teachers. Table 4 shows the results for this analysis where each row corresponds to a teacher and columns show the robust accuracy of both teacher and the robustified student on the ImagNet-C dataset. In majority cases difference in student accuracy between a given row and the row just upper it, is higher as compared to the same quantity but for the teacher column. This shows that when we switch to a better teacher (*i.e.* teacher accuracy gets increased) then the increment in student accuracy, in majority cases, is even more.

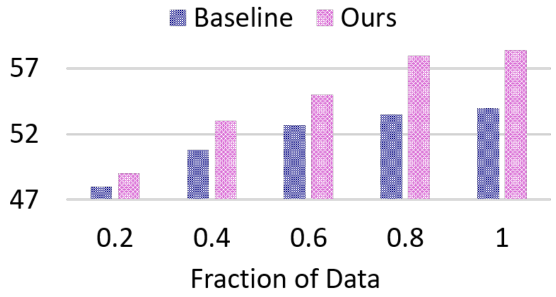

Figure 1. Ablation on amount of distillation data v/s accuracy. The x-axis shows the fraction of augmented data used in distillation and the y-axis shows the accuracy achieved for each fraction by the APT baseline and our method on the ImageNet-C dataset. Here, ResNet-101 is used as the student network, updated using a ResNet-34 teacher, both being single modal.

3. Further results on Transfer Learning

We further analyze our proposed scheme for the transfer learning setup under more datasets used in the CLIP paper for classification. Specifically, we use the Cars [3], CIFAR-100 [4], Aircraft [5] and SUN-397 datasets [6] for this setup and compare our model against the transfer learning of the original CLIP model and its completely fine-tuned version under the Visual Evaluation setting. Figure 2 shows the results for this analysis where original denotes the visual encoder of the initial CLIP model without any tuning. It can be observed, similar to the main paper, that complete fine-tuning is not able to preserve the transfer learning or generalization capabilities of the model whereas our method preserves this important characterstic of a pretrained model.

4. Analyzing more Multi-Modalities

We further analyze our method for other popular pretrained Multi-Modal Networks namely LiT[8] and UniCL[7]. For LiT, we use the LiT-B/16B model and for UniCL we use the SWIN-T model. Figure 3 shows the results for various teacher student pairs under Visual Evaluation setting on ImageNet-C,R datastes and under the Zero Shot setting on ObjectNet,ObjectNet-C datasets. For both settings Unimodal ResNet-101 and ViT-Small are used as teachers. Each vertical bar for a given student on x-axis (LiT/UniCL) correspond to a particular teacher and y-axis denotes the accuracy. For Visual Evaluation, we use APT baseline (rows with Teacher set to None) and for Zero-Shot, original Zero-Shot model (without any fine-tuning) is used as the baseline.

		Distribution Shifts						Transfer Learning		Avg.
	IN	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	IN-C	Tiny-IN	Flowers	shifts
CLIP *ViT-L/14@336px*										
Inference Time Ablations										
Ours w/o KLD	84.7	78.3	89.3	65.6	72.8	79.3	63.1	85.2	98.7	74.7
Ours w/o $\mathcal{U}_{mc}$	83.9	77.7	88.4	64.9	71.5	77.6	62.2	85.2	98.7	73.7
Ours (max logit)	84.2	77.6	88.1	64.1	71.5	79.2	62.9	85.2	98.7	73.7
Knowledge Distillation Ablation										
Ours w/o K.D.	84.9	77.6	86.8	63.2	70.3	78.3	62.1	**85.4**	**98.8**	73.1
Ours	**85.4**	**79.1**	**89.9**	**65.8**	**73.2**	**80.9**	**64.9**	85.2	98.7	**75.6**

Table 1. **Visual Evaluations results : Accuracy.** Comparison with ablations of our method discussed in Sec. 1 on all the distribution shifts used in the paper along with transfer learning experiments for the CLIP *ViT-L/14@336px* model . The last column shows the average accuracy over all the shifts. The first row of numbers correspond to the version of our method without KL divergence term at inference time. Similarly, the row below it corresponds to our method but without uncertainty ($\mathcal{U}_{mc}$) term at inference time. The second last row corresponds to the variant of our method without any knowledge distillation and the last row correspond to our complete method using single modal ViTB/16 teacher.

		Distribution Shifts				
	IN	IN-R	IN-Sketch	ObjectNet	IN-A	IN-C
w/o KLD	0.89	0.82	0.71	0.82	0.79	0.83
w/o $\mathcal{U}_{mc}$	0.71	0.69	0.72	0.61	0.63	0.71
Ours	**0.95**	**0.94**	**0.89**	**0.84**	**0.92**	**0.94**

Table 2. **Visual Evaluations results :** $F_{correct}$. Analysis of our method and its ablations (discussed in Sec. 1) using the $F_{correct}$ (fraction of examples for which correct head is selected) metric described in Sec. 1, for the CLIP *ViT-L/14@336px* model. Here, *w/o KLD* denotes the ablation without the KL divergence at inference time and similarly *w/o* $\mathcal{U}_{mc}$ denotes the ablation without the uncertainty term ($\mathcal{U}_{mc}$) during inference. For our method single modal ViT-B/16 is used as teacher.

Student	Teacher	*Ours w/o K.D.*	Ours
RN-101	RN-34	53.8	55.3
CLIP RN-101	RN-101	54.9	56.7
CLIP ViT-B/16	ViT-S	55.1	57.4
CLIP ViT-B/32	ViT-S	54.5	57.1

Table 3. **Knowledge Distillation Ablation : Accuracy**. Comparison with the *w/o K.D.* (without knowledge distillation) ablation of our method (refer Sec. 1) using both single and multi-modal (CLIP) networks as students and single modal teachers, under a Visual Evaluation setup on the ImageNet-C dataset.

It can be observed that our method again improves the accuracy over the APT baseline by around 2.5% (average) on the ImageNet-C dataset and by around 1.9% (average) on the ImageNet-R dataset under the Visual Evaluation setting. Similarly for Zero Shot setting, it improves accuracy by around 2.5% (average) over the baseline on the ObjectNet-C dataset. This shows that our method generalizes well to other pretrained multi-modalities as well.

CLIP *ViT-B/16* Student (149M)			
	Params	Teacher	Student
RN-34	11M	55.7	54.3
RN-50	23M	57.6	56.9
RN-101	45M	58.7	57.2
ViT-Small	48M	60.9	57.4
RN-151	60M	62.4	59.7
ViT-Base	88M	63.2	62.5

Table 4. **Visual Evaluation Results : Accuracy**. Analyzing how robust accuracy of CLIP ViT-B/16 student changes with increasing robust accuracy of the teacher. Here, each row corresponds to a robustified teacher with architecture given in the first column. The teacher column shows the accuracy of this robustified teacher on the ImageNet-C data. Similarly, each student column element shows the accuracy of the fixed student on the ImageNet-C data, when the teacher corresponding to its row is used in our method to distill knowledge.

5. Analysis on ImageNet-P

We further analyze robustness using the mean Flipping Rate (mFR)[2] for the ImageNet-P dataset for various CLIP models (RN-50,101 and ViT-B/16) under the Visual Evaluation setting.
We begin by first analyzing the robustness of the teacher models, robustified using the scheme described in the main paper (using AugMix+DeepAugment). Table 6 shows this analysis, comparing the mFR metric for these single-modal teacher networks, with and without (Naive) applying the robustification scheme. It can be observed that their performance is significantly improved after robustification. The best performance is shown by the ViT-Small model. Even though the model has comparable parameters to ResNet-101, still a significant difference in the performance highlights its robust learning scheme. Given the robustified

	RN-101	RN-151	ViT-S	ViT-B16	RN-50C	RN-101C	ViT-B16C	ViT-L14C	LiT-B16/B
Tuned	2.1M	2.6M	2.3M	3.1M	3.9M	4.1M	4.3M	5.6M	4.4M
Total	45M	60M	48M	88M	102M	119M	149M	450M	195M

Table 5. Number of parameters tuned and total number of parameters for all the students used. Here, each column corresponds to a student network architecture used in this work. First row shows the number of parameters for a given column as student, when our method is applied to update this student. The last row shows the total parameter count of this student.

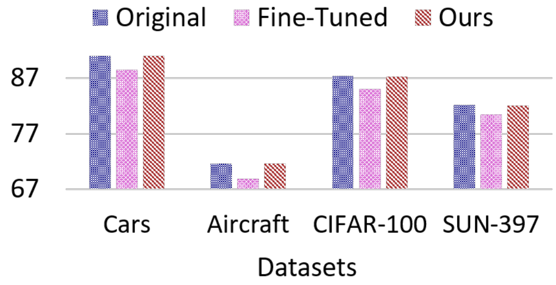

Figure 2. Transfer Learning under Visual Evaluation setup. The x-axis consist of various datasets and y-axis shows the accuracy on each of these datasets under the transfer learning setup. This analysis is done for original pre-trained CLIP *ViT-L/14@333px* model (Original), its ImageNet fine-tuned version (Fine-tuned) and after it has been updated using our approach with a ViT-B/16 teacher (Ours).

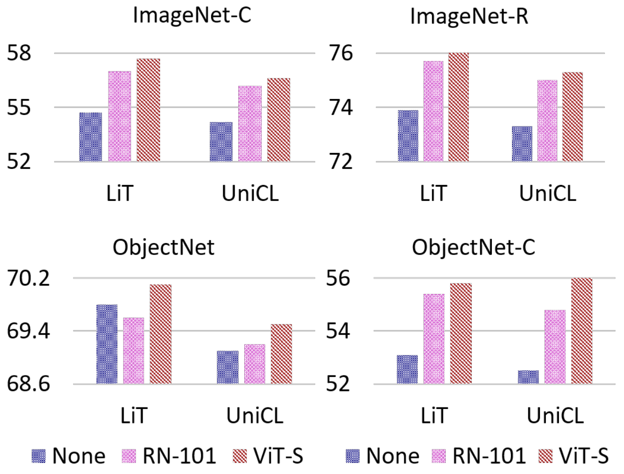

Figure 3. **Visual Evaluation Results : Multi-Modalities**. y-axis : Accuracy, x-axis : Multi-Modal Student, vertical bar : Teacher.This figure analyses our method on the ImageNet-C,R and ObjectNet,ObjectNet-C datasets when LiT-B/16B, UniCL SWIN-T multi-modal networks are used as students (x-axis) and are tuned using single modal networks (RN-101, ViT-S) as teacher (vertical bar). Vertical bar labelled None correspond to APT baseline. It can be observed that out method provides significant gains over the APT baseline, especially when ViT-S model is used as the teacher.

teachers, we now evaluate students tuned using our method on this dataset. Figure 4 shows the results for the CLIP model students using our method and also for the APT baseline. It can be observed that our method is able to reduce mFR by upto 3% and greater than 2% for the cases with ViT-Small as the teacher. Even thouh ViT-Small and RN-101 have similar number of parameters, still a significant difference is observed for CLIP RN-50 and CLIP ViT-B/16 model showing that ViT transfers the robustness more efficiently w.r.t. this dataset.

Arch.	Naive	AugMix+DA
RN-34	59.8	41.6
RN-50	57.1	37.5
RN-101	52.7	34.6
ViT-Small	49.2	33.4

Table 6. Comparison of mean Flipping Rate (mFR) on ImageNet-P dataset to analyze robustness induced by applying AugMix+DA for tuning the relatively small teacher models.

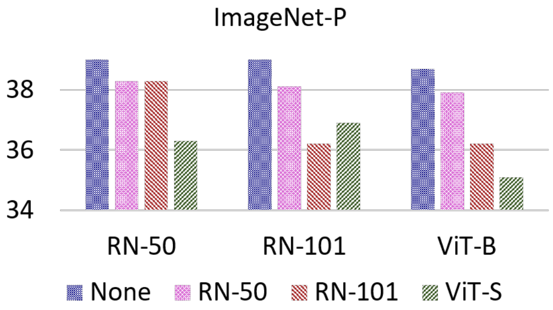

Figure 4. **Visual Evaluation Results : mFR**. y-axis : mFR, x-axis : CLIP Student Network architectures, vertical bar : Teacher. This figure analyses our method on the ImageNet-P dataset when various CLIP model architectures are used as students (x-axis) and tuned using various single modal networks as teacher (vertical bar). Vertical bar labelled None correspond to APT baseline. It can be observed that out method provides significant gains over the APT baseline, especially when ViT-S model is used as the teacher.

6. Parameters Tuned

We provide the number of parameters tuned for each network used as student and updated using our approach along with their total parameter count in table 5. Unless mentioned the number of parameters tuned for a given student in all of the experiments corresponds to the value provided in this table.

7. Implementation Details

The network tuning using our algorithm involves updating a portion (refer table 5) of the complete architecture starting from end using a learning rate of $1e-3$ and a batch size of 256 using an Adam optimizer. For tuning our method, unless mentioned, we use the half of the complete augmented data (a=0.5), generated by DeepAugment+Augmix. The tuning is carried out for 500 epochs. Robustifying the teacher involves the same pipeline and hyper-parameters as proposed in the DeepAugment paper[1]. Also a dropout is applied while training with probability set to 0.25. The last one-fifth of the tuning portion corresponds to clean/unclean and combined heads. At the inference time, $N = 10$ samples are drawn per head with dropout activated to estimate uncertainty term ($\mathcal{U}_{mc}$). Not, this requires multiple passes only through dropout activated layers, not the complete model.

References

[1] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 5

[2] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[5] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2

[6] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2

[7] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2

[8] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2