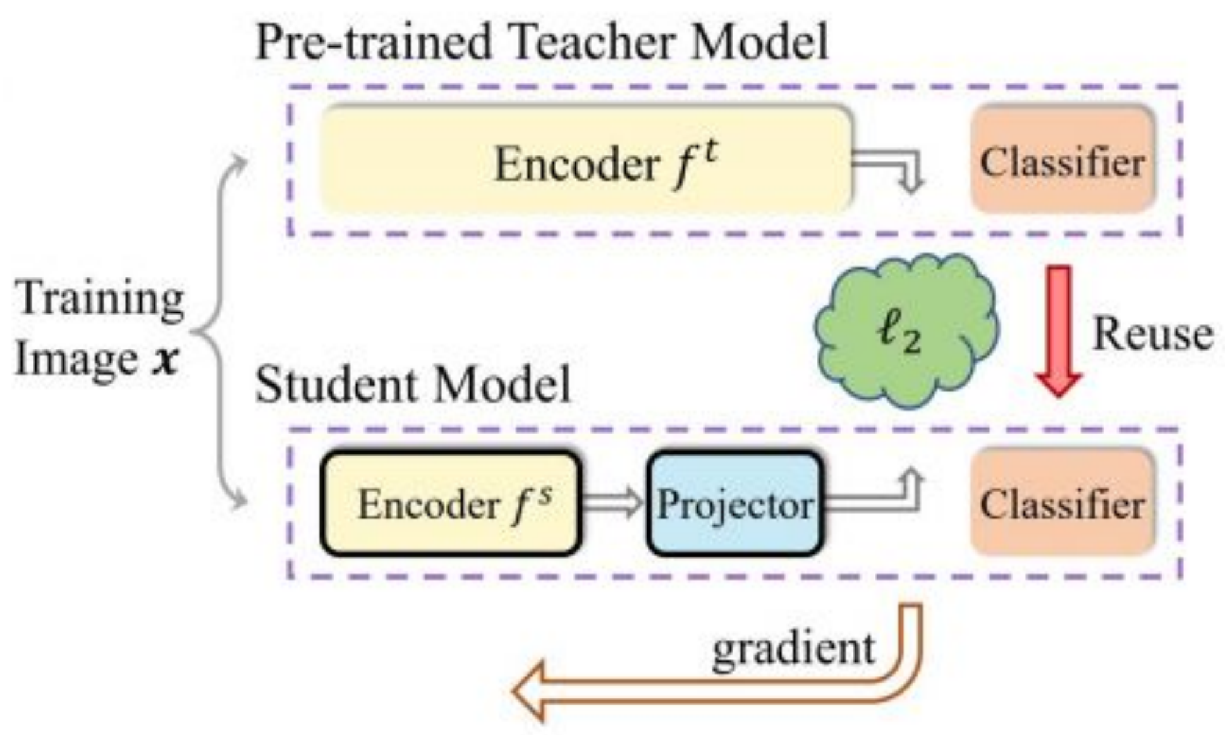
Knowledge Distillation with the Reused Teacher Classifier

Wut Hmone Hnin Hlaing, Steffen Röber, Apurv Kumar, Brhanu Atsbaha, Abin Baby

Introduction

Knowledge distillation aims to compress a powerful yet cumbersome teacher model into a lightweight student model without much sacrifice of performance. Our study explores knowledge distillation using simKD model as a baseline with diverse teacher-student architectures with varying learning rates, loss functions, multiple projectors, labels, and unbiased projectors to enhance feature alignment and model performance.



SimKD model

- Loss Functions - L2 Loss (SimKD), KL divergence
- Learning rates - 0.001, 0.005, 0.01, 0.05, 0.1
- Multiple Projectors (baised)
- Using labels
- All Layer projectors (unbaised)

Different loss experiments and learning rates

L2 Loss, KL Divergence

For our experiments, we primarily used **L2 loss** and **KL Divergence**. L2 loss, also known as Mean Squared Error (MSE), measures the squared differences between predicted and actual values, making it suitable for tasks requiring accurate regression or prediction. On the other hand, KL Divergence (Kullback-Leibler Divergence) quantifies the difference between two probability distributions, often used in probabilistic models and machine learning to ensure alignment between predicted and target distributions.

$$\text{L2 Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

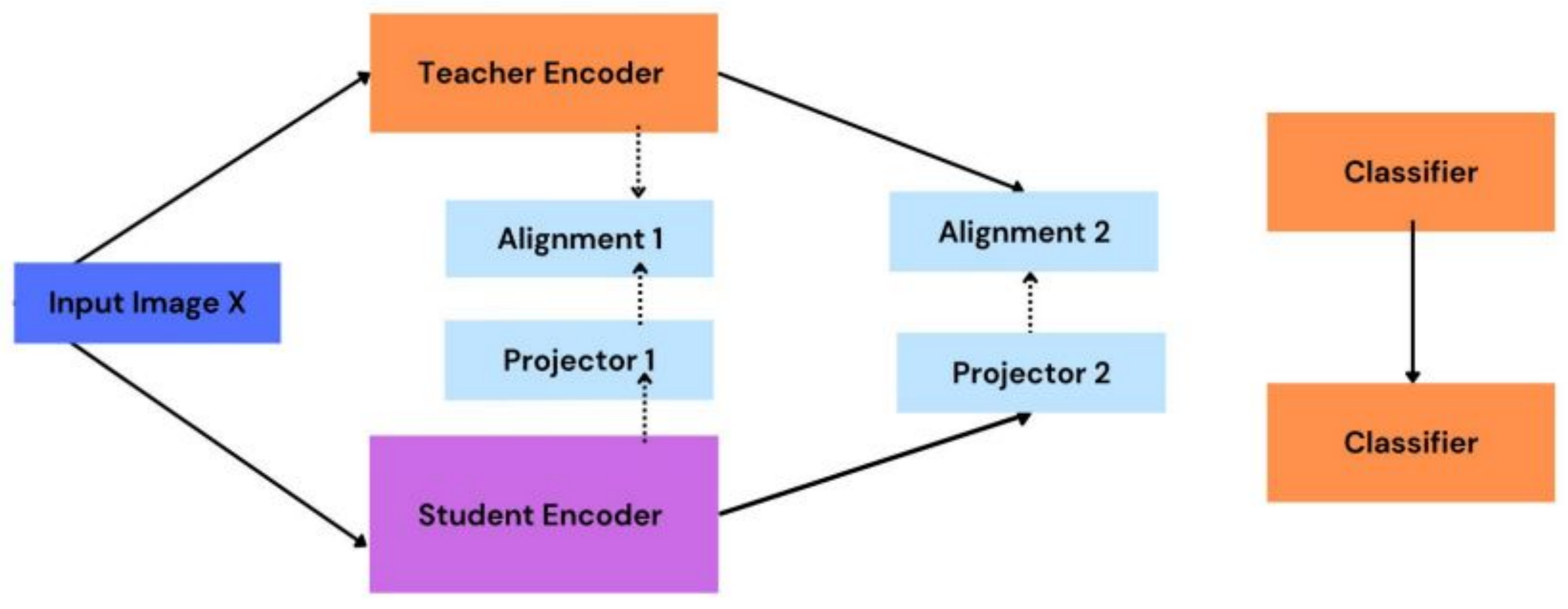
L2 Loss

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$$

KL Divergence

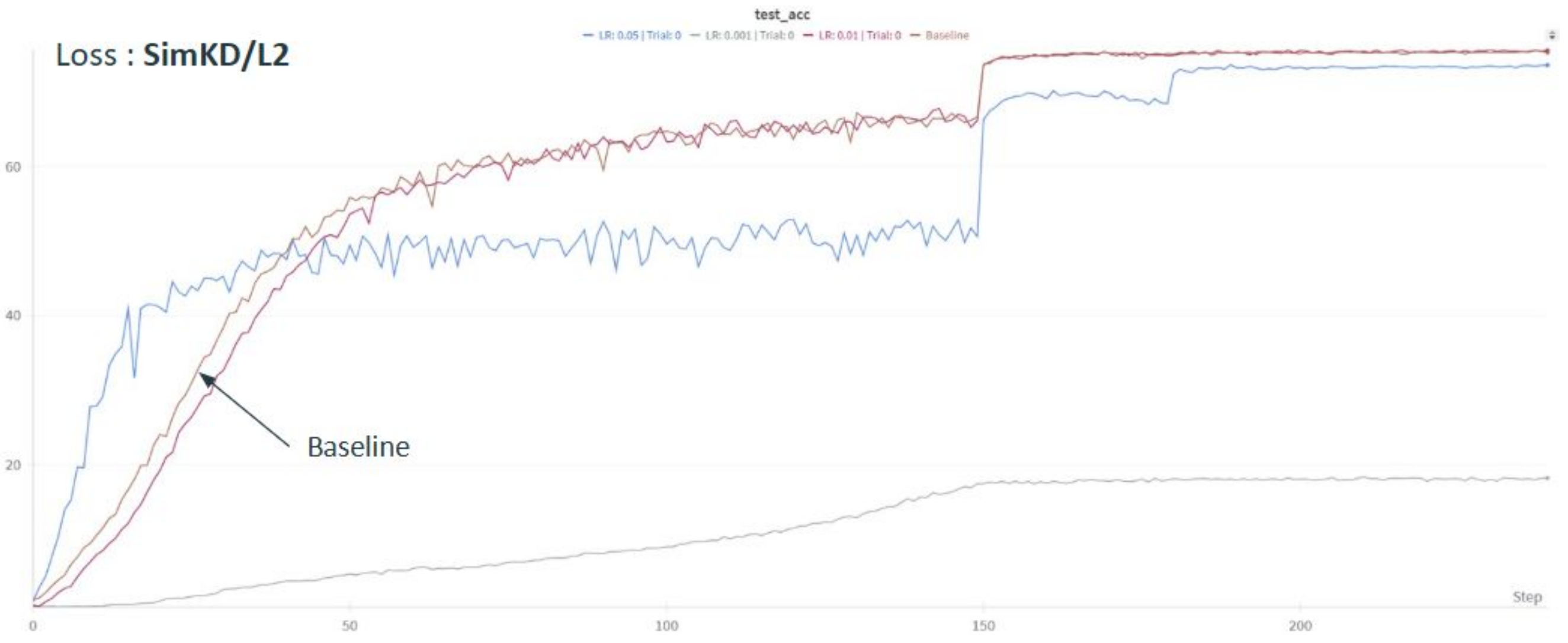
Multiple projectors (baised)

One after 2/3 of the model's layers and one at the end



Multiple projectors

In a multiple projectors experiment employing the Teacher architecture ResNet-32×4 and MobileNetV2_1_0, two loss functions, L2 loss and KL Divergence, were evaluated across learning rates of 0.01, 0.05, and 0.001. The L2 loss achieved a maximum accuracy of 75.52 ± 0.18 at a learning rate of 0.01, surpassing the baseline accuracy of 75.43. Conversely, KL Divergence attained a maximum accuracy of 66.88 ± 3.78, also at a learning rate of 0.01. These results underscore the superior performance of the L2 loss function in this experimental setup, demonstrating its potential for more effective knowledge distillation in this context.



Multiple projector using L2 loss

Using Labels

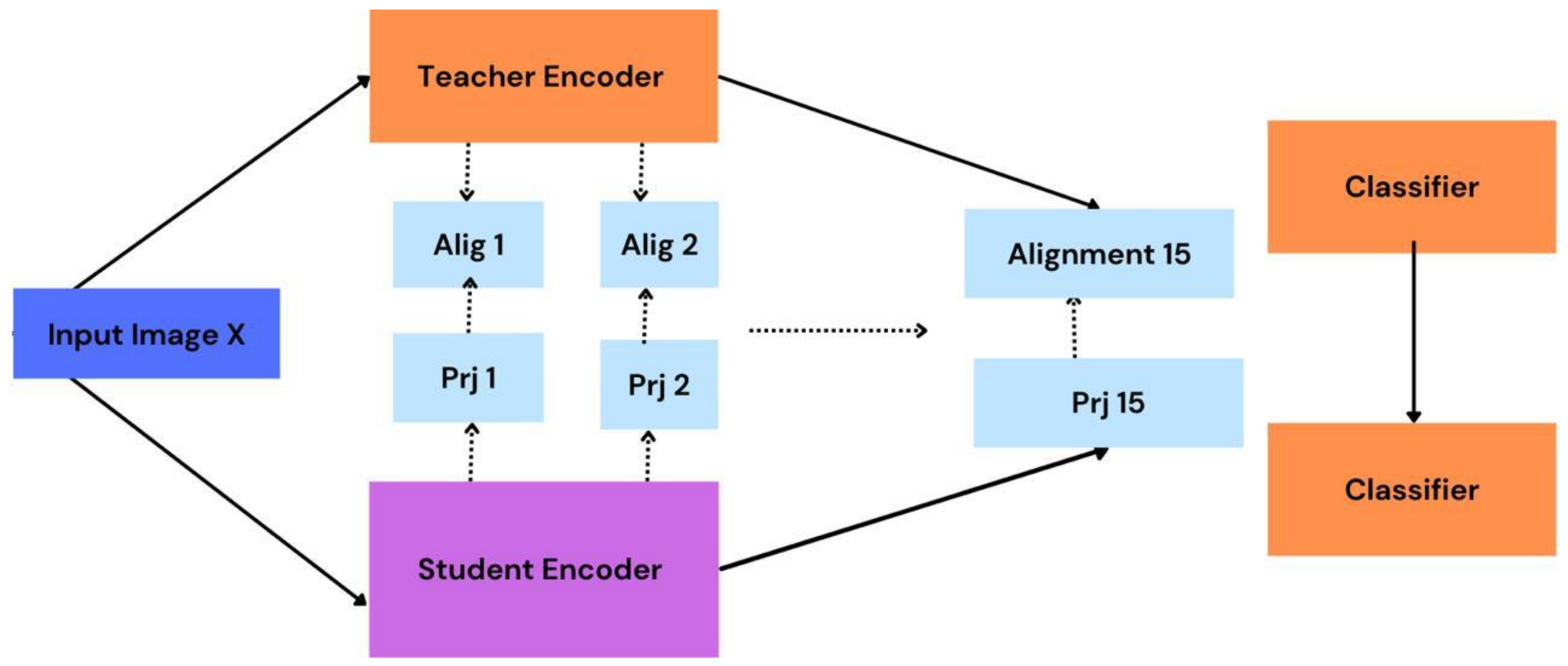
Let the student learn, only if the teacher made a correct prediction

In an experiment utilizing labels for training, the Teacher architecture ResNet-32×4 and MobileNetV2_1_0 were employed with L2 loss and KL Divergence across learning rates of 0.01, 0.05, and 0.001. L2 loss achieved a maximum accuracy of 75.41 ± 0.48 at a learning rate of 0.01, which is on par with the baseline accuracy of 75.43, demonstrating its effectiveness in maintaining performance consistency. In contrast, KL Divergence reached a maximum accuracy of 66.87 ± 3.78 at the same learning rate, indicating a less stable performance compared to L2 loss in this experimental setup.

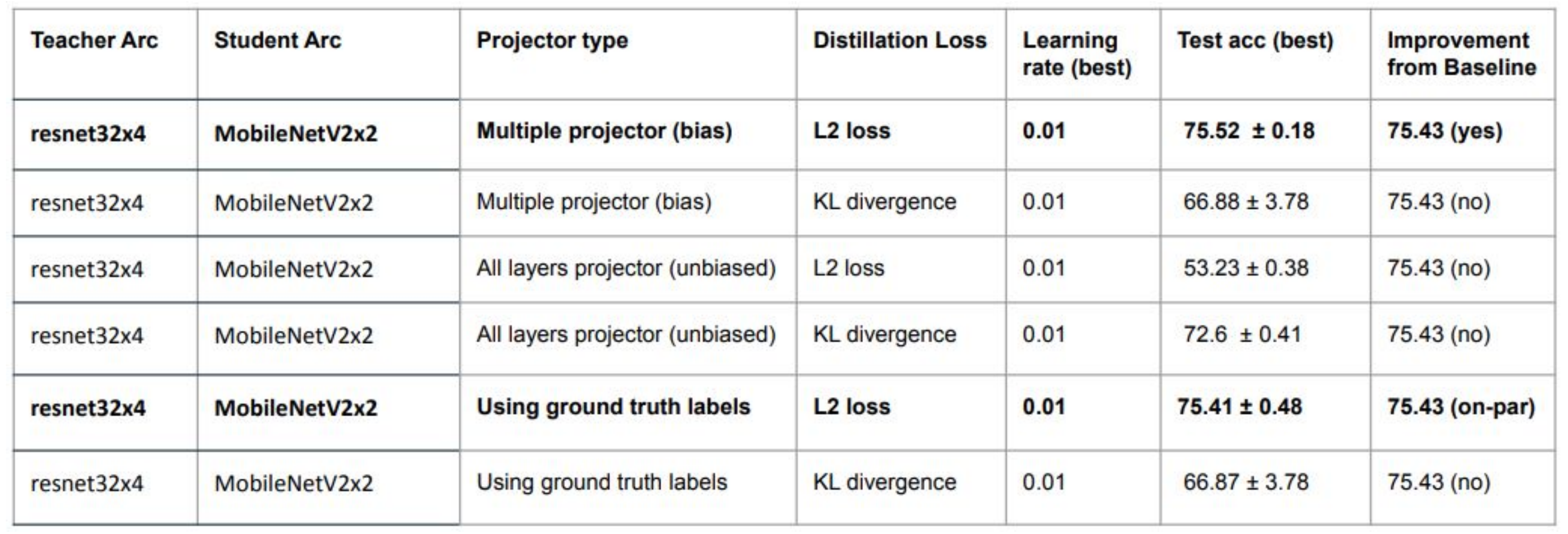
All Layer Projector (unbaised)

Projectors at all layers of the student model

In an experiment incorporating projectors after every layer to ensure continuous and detailed alignment during training, the Teacher architecture ResNet-32×4 and MobileNetV2_1_0 were utilized with L2 loss and KL Divergence at learning rates of 0.01, 0.05, and 0.001. Among the configurations tested, L2 loss achieved a maximum accuracy of 53.23 ± 0.38 at a learning rate of 0.01, while KL Divergence demonstrated superior performance, reaching a maximum accuracy of 72.6 ± 0.41 at the same learning rate. These findings highlight the effectiveness of KL Divergence in leveraging the layered projector alignment approach compared to L2 loss under similar conditions.



All layer projectors



Teacher Arc	Student Arc	Projector type	Distillation Loss	Learning rate (best)	Test acc (best)	Improvement from Baseline
resnet32x4	**MobileNetV2x2**	**Multiple projector (bias)**	**L2 loss**	**0.01**	**75.52 ± 0.18**	**75.43 (yes)**
resnet32x4	MobileNetV2x2	Multiple projector (bias)	KL divergence	0.01	66.88 ± 3.78	75.43 (no)
resnet32x4	MobileNetV2x2	All layers projector (unbiased)	L2 loss	0.01	53.23 ± 0.38	75.43 (no)
resnet32x4	MobileNetV2x2	All layers projector (unbiased)	KL divergence	0.01	72.6 ± 0.41	75.43 (no)
resnet32x4	**MobileNetV2x2**	**Using ground truth labels**	**L2 loss**	**0.01**	**75.41 ± 0.48**	**75.43 (on-par)**
resnet32x4	MobileNetV2x2	Using ground truth labels	KL divergence	0.01	66.87 ± 3.78	75.43 (no)

Overview of all the experiments

References

Chen, D., Mei, J.-P., Zhang, H., Wang, C., Feng, Y., & Chen, C. (2024). Knowledge Distillation with the Reused Teacher Classifier.

Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D.Lawrence, and Zhenwen Dai (2019). Variational information distillation for knowledge transfer

Hailin Zhang, Defang Chen, and Can Wang (2021). Confidence aware multi-teacher knowledge distillation.