Transferring Inductive Bias through Leveled Knowledge Distillation

Clement Wong[1], Derek Fermin[2], Gretchen Zheng[3], Jeongsoo Park[3], Zilin Wang[3]

[1]Mechanical Engineering. [2]Robotics. [3]Electrical Engineering and Computer Science

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

1. Motivation & Objective

Motivation:

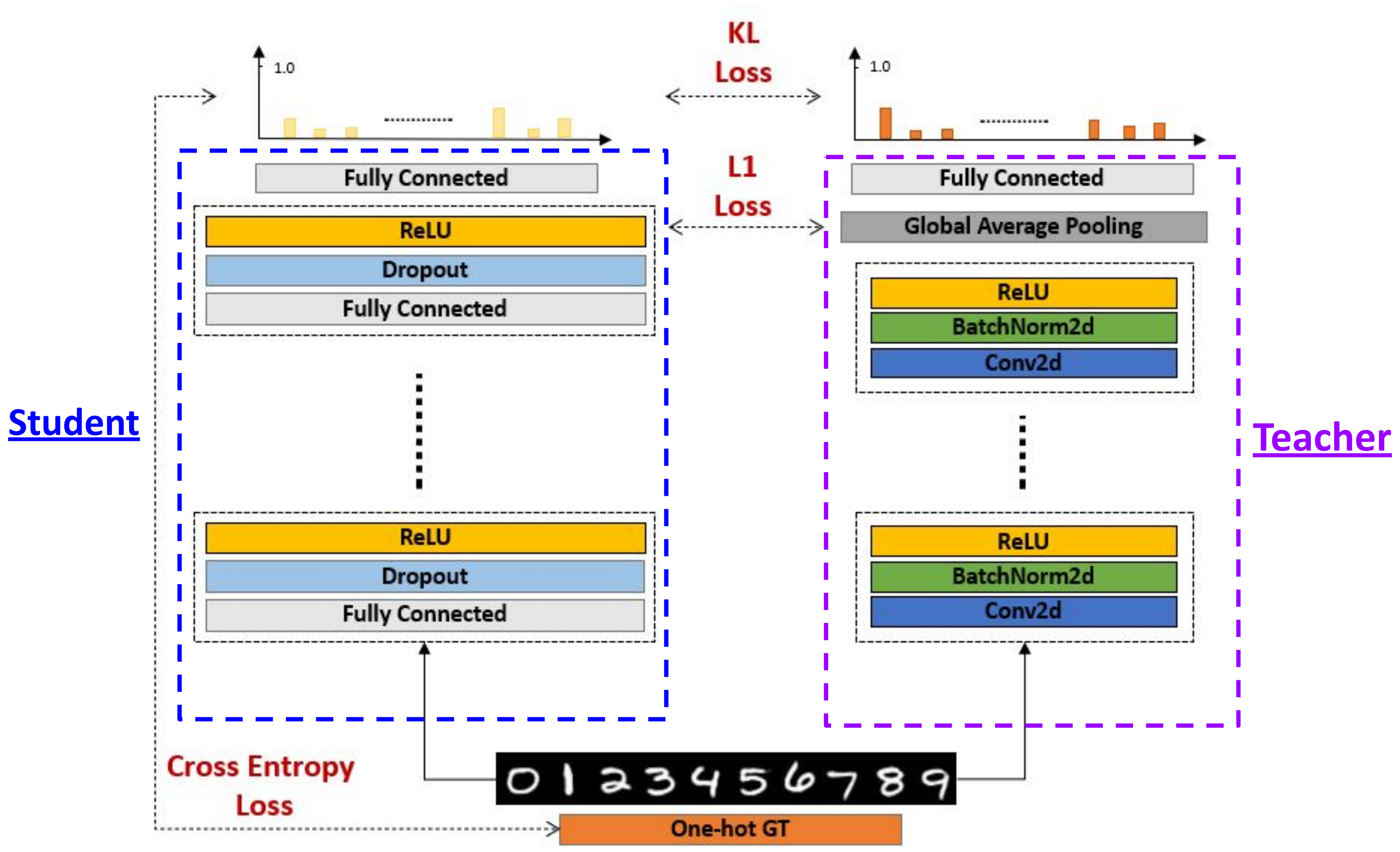
- Using a computationally expensive model architecture on a resource-constrained edge device is not feasible
- But the inductive bias of those models, such as CNN's translational equivariance, is sometimes necessary
- Knowledge Distillation (KD) is known to transfer knowledge from a more sophisticated neural network to a simpler one

Objective:

- Explore if Knowledge Distillation transfers the inductive bias from the teacher model to a student model with a simpler architecture
- If so, try to improve the transfer of the inductive bias by improving the Knowledge Distillation scheme

2. Knowledge Distillation

- Use a sophisticated model with inherent inductive bias as "teacher"
- Design a simple model without any known inductive bias as "student"
- Try to match the prediction of "student" model with the "teacher" model's predictions
- Also we experiment on matching the intermediate features of both models



3. Method

Loss functions for Knowledge Distillation:

- Cross Entropy Loss (with the ground truth): $$L_{CE} = -\log\left(\frac{\exp(s_{y_i})}{\sum_j \exp(s_j)}\right)$$
- KL Divergence Loss (matching the predicted probability distributions between student and teacher): $$L_{KL} = KL((\text{softmax}(\frac{\psi_{student}(\mathbf{X})}{\tau}), \text{softmax}(\frac{\psi_{teacher}(\mathbf{X})}{\tau}))$$
- L1 Loss: $$L_{FM} = \sum_{i=1}^{n} |y - \hat{y}_i|$$
- Proposed Loss (α, τ, s_{KL}, s_{FM} are hyperparameters)

$$L = \alpha L_{CE} + (1-\alpha)(s_{KL}\tau^2 L_{KL} + s_{FM} L_{FM})$$

Evaluation:

- Train MLP on MNIST and CIFAR-10
- Train MLP on MNIST and CIFAR-10 with Knowledge Distillation
- Evaluate the both MLPs' accuracy on translated MNIST and CIFAR-10
- Measure the difference in accuracy to see if CNN's inductive bias (translational equivariance) has been transferred

4. Experiments

Training settings:

- Dataset: MNIST, CIFAR-10
- Optimizer: AdamW; weight decay: 0.001; learning rate: 0.001

Model architectures:

- CNN_1: 4 convolutional layers with batch normalization and ReLU activation; # of channels: [32, 64, 128, 256]; global average pooling after the last convolutional layer
- CNN_2: 4 convolutional layers with batch normalization and ReLU activation; # of channels: [64, 64, 128, 128]; global average pooling after the last convolutional layer
- MLP_1: hidden units: [2000, 1000, 100]; dropout rate: 0.3; ReLU activation
- MLP_2: hidden units: [1024, 512, 256, 128]; dropout rate: 0.3; ReLU activation

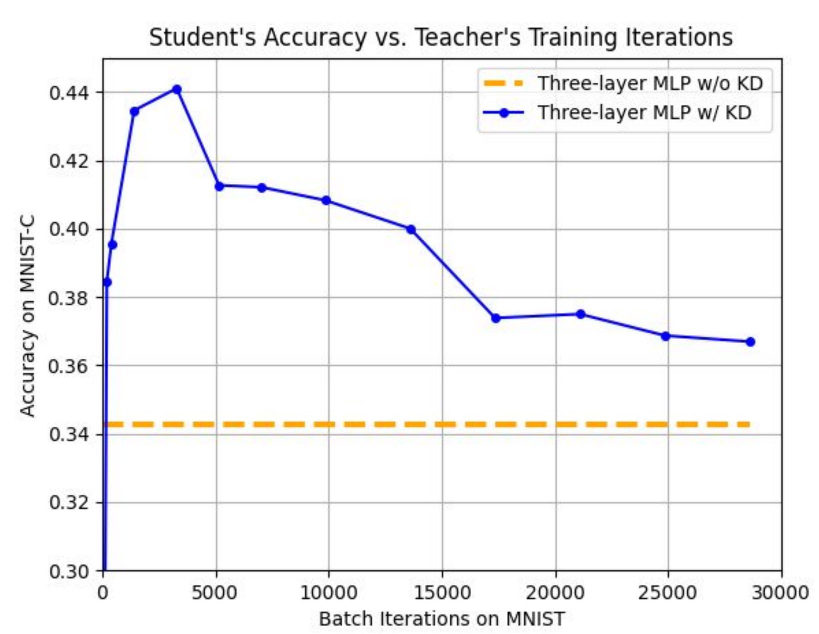
Experiments:

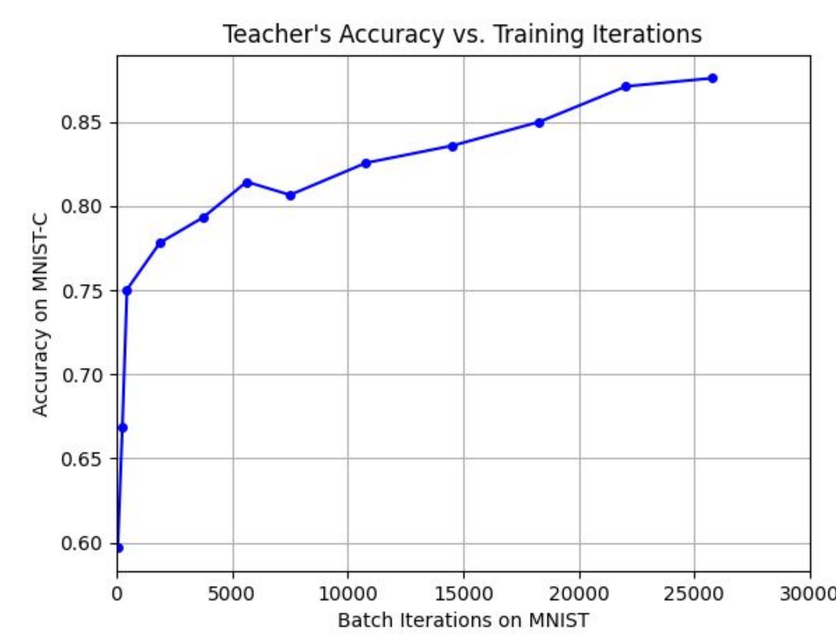
- Experiment 1:
  - Use different snapshots of CNN_1 as teachers using normal Knowledge Distillation approach as in [2].
    - Teacher model is CNN_1 and student model is MLP_1
  - Measure the performance of student model per trained batch iterations of the teacher
- Experiment 2:
  - Build on the result of experiment 1; Try to improve the KD scheme by implementing feature matching
  - Compare weakly-trained and strongly-trained version of CNN_2
    - Teacher model is CNN_2 and student model is MLP_2
    - Compare weak teacher versus strong teacher + label smoothing
  - Try and compare different feature matching settings
  - Investigate the contribution of ground truth cross entropy loss
    - Compare the results from different α

5. Results

Result 1: Weaker Teacher → Better Performance

- The vanilla KD was able to increase the robustness of the student trained on MNIST dataset
- With a batch size of 128, we take snapshots of the CNN_1 model every 50 batch iterations in the first 500 batch iterations; after that, we increase to every 500 batch iterations.
- CNN_1 does not seem to overfit since it's accuracy increases monotonically as number of batch iterations gets higher
- The CNN_1 trained for ~2500 batch iterations (~5 epochs) produced the most robust student performance
- Weaker teachers around ~2500 batch iterations show better student performance than stronger teachers with 25000+ batch iterations

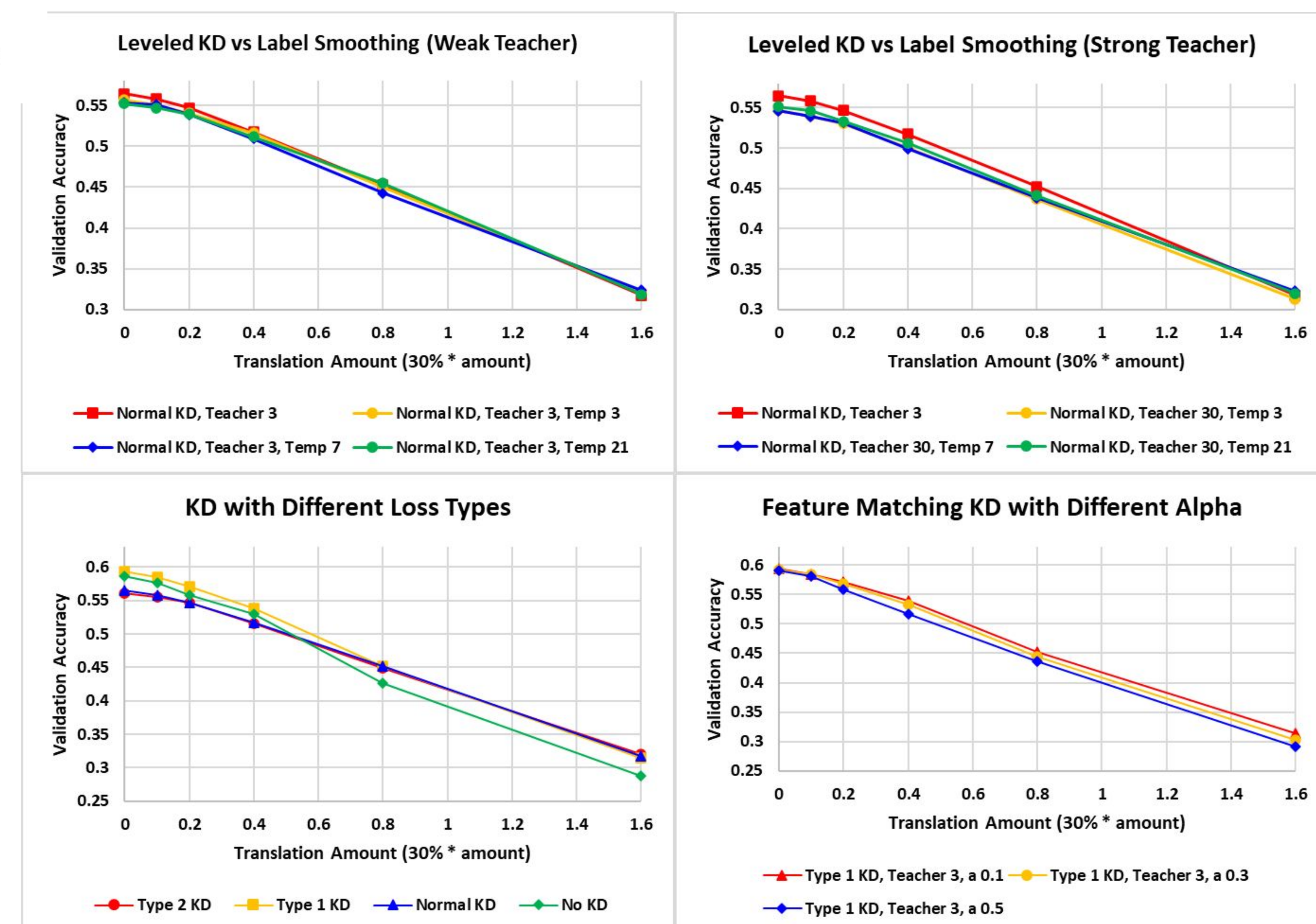




Result 2: Feature Matching + Weaker Teacher enhances inductive bias transfer & improves accuracy

- Weakly-trained vs strongly-trained teacher
  - Weaker teacher replaces label smoothing and offers better performance
    - Leveled KD (KD using weaker teachers) works better than stronger teacher with label smoothing
- Different feature matching settings
  - No KD: Ground Truth label matching (s_{KL}, s_{FM}=0)
  - Normal KD: Ground Truth + Teacher's Prediction matching (s_{FM}=0)
  - Type 1 KD: Ground Truth + Teacher's Feature Space matching (s_{KL}=0)
  - Type 2 KD: Ground Truth + Teacher's Feature Space + Teacher's Prediction matching

  → *Type 1 KD seems to perform the best in terms of inductive bias transfer and highest validation accuracy*
- Different α experiment
  - α controls the ratio between ground truth cross entropy loss and KD loss (teacher's prediction, feature space)
    - Bigger α means the portion of the ground truth cross entropy loss gets bigger
  - As α gets larger, the trained student model becomes less robust to translation → less inductive bias transfer from the teacher



6. Conclusions

- Our results indicate that Knowledge Distillation can help a simple student model to learn the more sophisticated teacher model's inductive bias
- *Weaker Teacher:* We found out that we can use weaker teacher as a replacement for label smoothing. Our experiment shows that weak teacher actually enhances inductive bias transfer versus the strong teacher with label smoothing
- *Feature Matching:* We demonstrated that feature matching improves the transfer of inductive bias by a noticeable margin while also boosting the performance of student model

7. References

[1] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network
[2] S. Abnar et al, Transferring inductive biases through knowledge distillation.
[3] R. Muller, S. Kornblith, and G. Hinton. When Does Label Smoothing Help?