Samsung Research

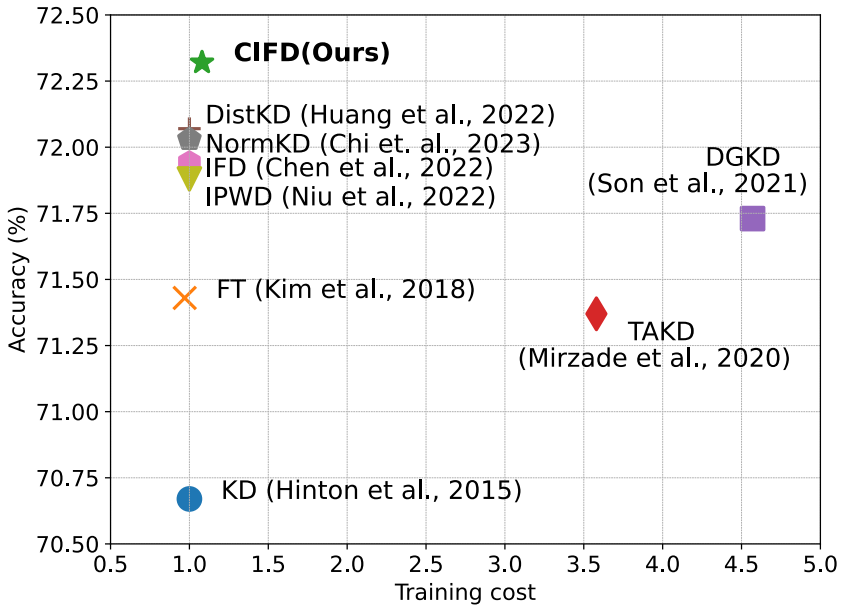
CIFD: Controlled Information Flow to Enhance Knowledge Distillation

Yashas Malur Saidutta, Rakshith S Srinivasa, Jaejin Cho, Ching-Hua Lee, Chouchang Yang, Yilin Shen, Hongxia Jin

12/12/24

Problem Statement & Prior work

- Knowledge Distillation: Transfer knowledge from large teacher model to small student model
- Problem: Transferring information from teacher to student when model capacity gap is large.
- Prior works:
  - Use “Teacher Assistants” (TA) to help facilitate transfer
    - Teacher Assistant: Intermediate models of size between teacher and student
  - Use novel loss functions to facilitate better transfer
- Drawbacks:
  - TAs are expensive to train
  - Loss functions do not account for the teacher-student capacity gap
- Proposed:
  - Mimic TAs by reusing low-level features of teacher model



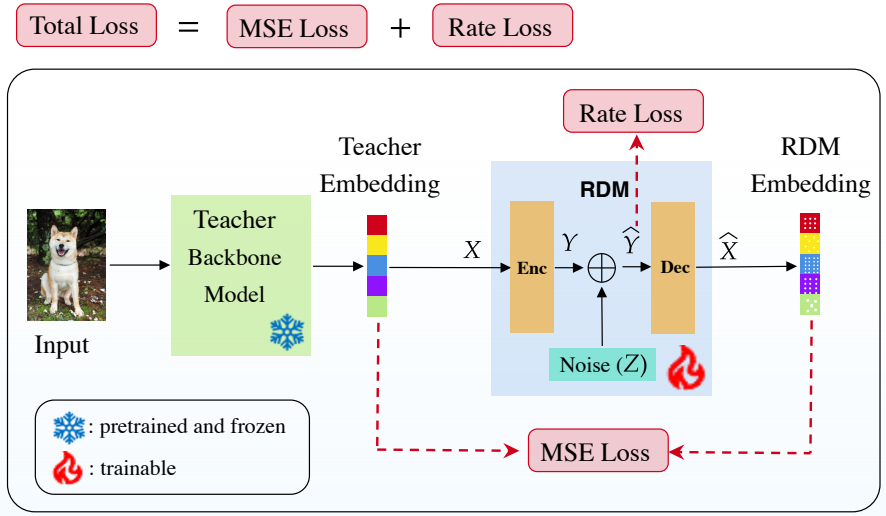
Total training cost of training a ResNet18 student from a ResNet34 teacher.

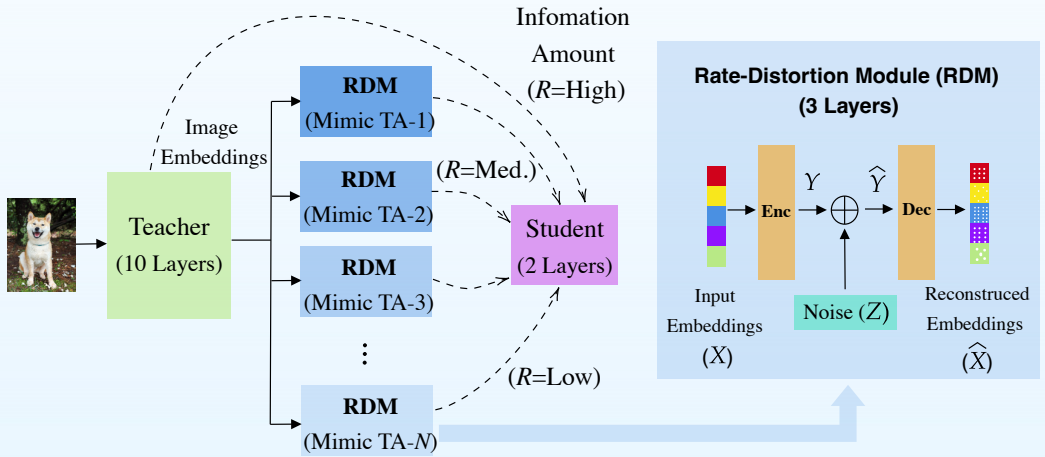
Presentation Outline

- Introduction to the problem
- Contributions
  - RDMs (Rate-Distortion Modules) to mimic Teacher Assistants
  - IBM (Information Bottleneck Modules) to improve performance
- Experimental results
  - Classification models on ImageNet-1k
  - Distillation of CLIP like models
  - Results on large teacher-student capacity gaps
- Conclusion

Contribution 1: Rate Distortion Module to mimic TAs

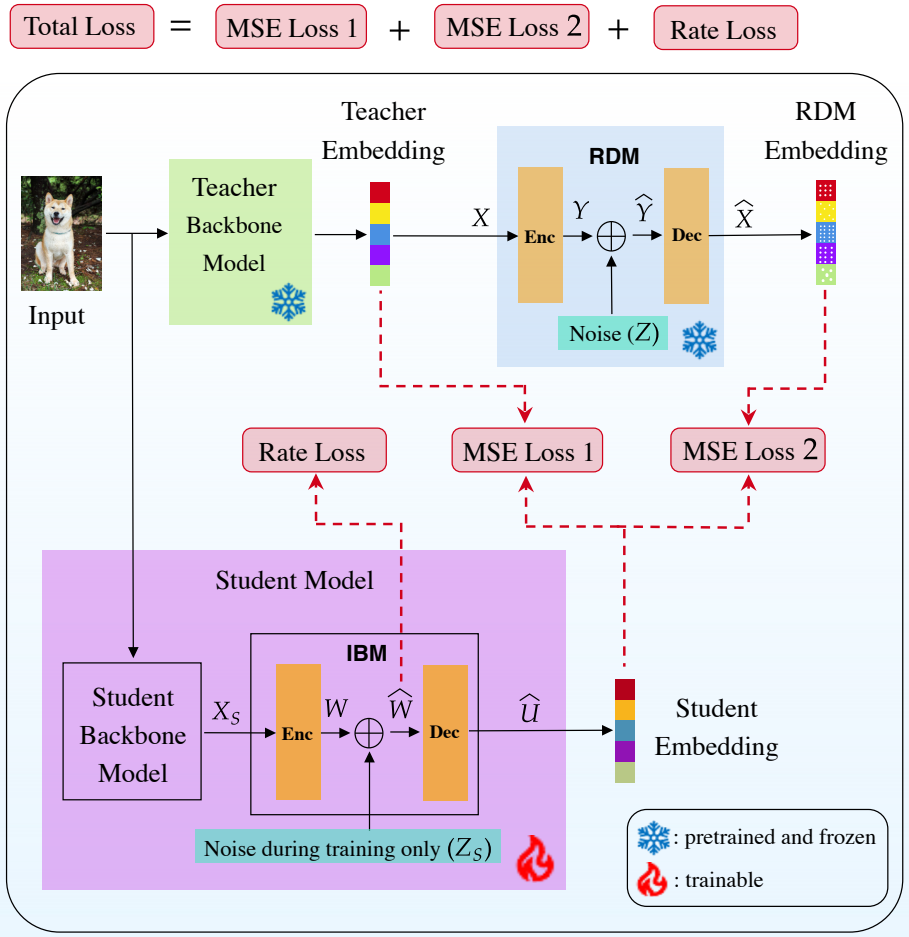
- RDMs mimic TAs by passing the teacher embedding through a “compression channel”
- RDM loss has two components
  - Reconstruction loss
  - Rate loss
- Why does RDM work?
  - Smaller TAs learn simpler concepts because model capacity is limited.
  - In RDMs, more the compression, simpler concepts are transmitted across noisy channel.
- More compression implies more distortion, smaller the TA that the RDM is mimicking
- Pros of RDMs:
  - No relearning of low level features from input.
  - RDMs can be trained in parallel and hence multiple can be trained and used.





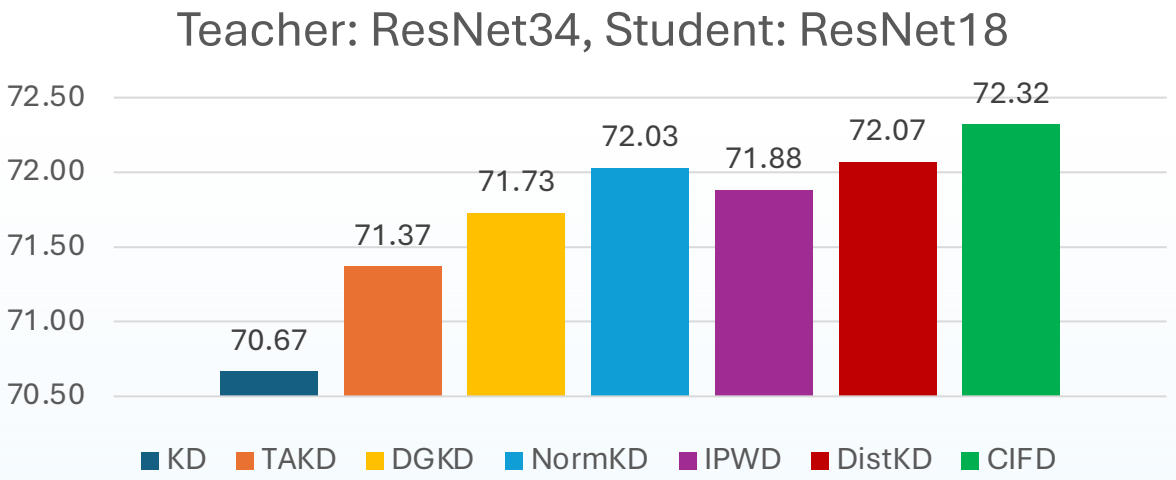
Contribution 2: Information Bottleneck Module

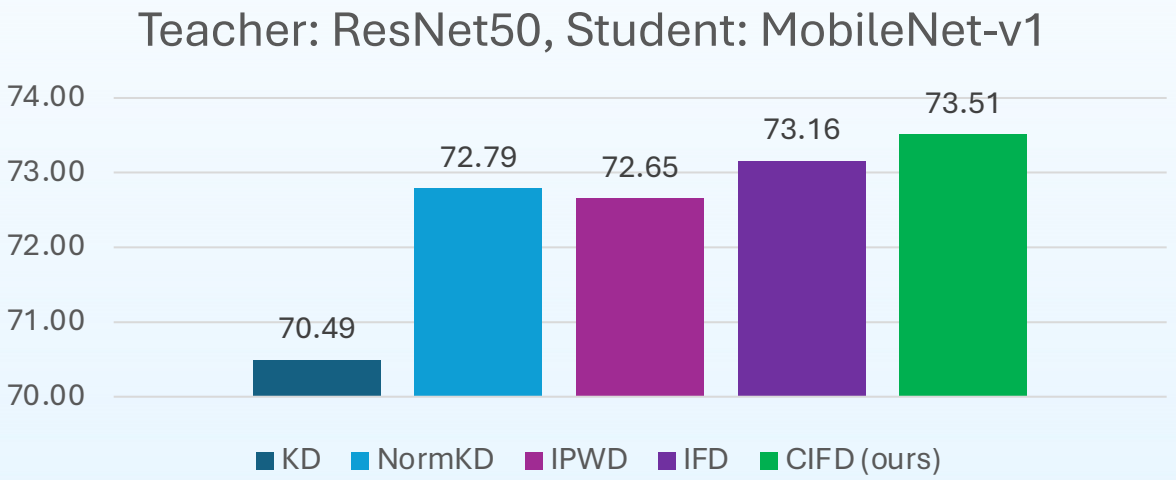
- We introduce an Information Bottleneck Module (IBM) in student model.
  - Designed from the Information Bottleneck Principle
- IBM is removed during inference.
- IBM controls the amount of feedback from RDMs.
  - Crucial to prevent overfitting with more RDMs.
- IBM can also be used independently.
- IBM can be shown to be an upper-bound on IBP like Masked Image Modeling.
  - Connections between MIM and IBP explored in Sec. 3.2 and Appendix C of paper.
  - IBP improves generalization, hence, MIM also helps improve generalization.



Results on ImageNet-1k classification

- Dataset: ImageNet-1k
- Two teacher-student combos
  - Teacher and student with same arch style
    - Teacher: ResNet 34
    - Student: ResNet 18
  - Teacher and student with different arch style
    - Teacher: ResNet 50
    - Student: MobileNet V1

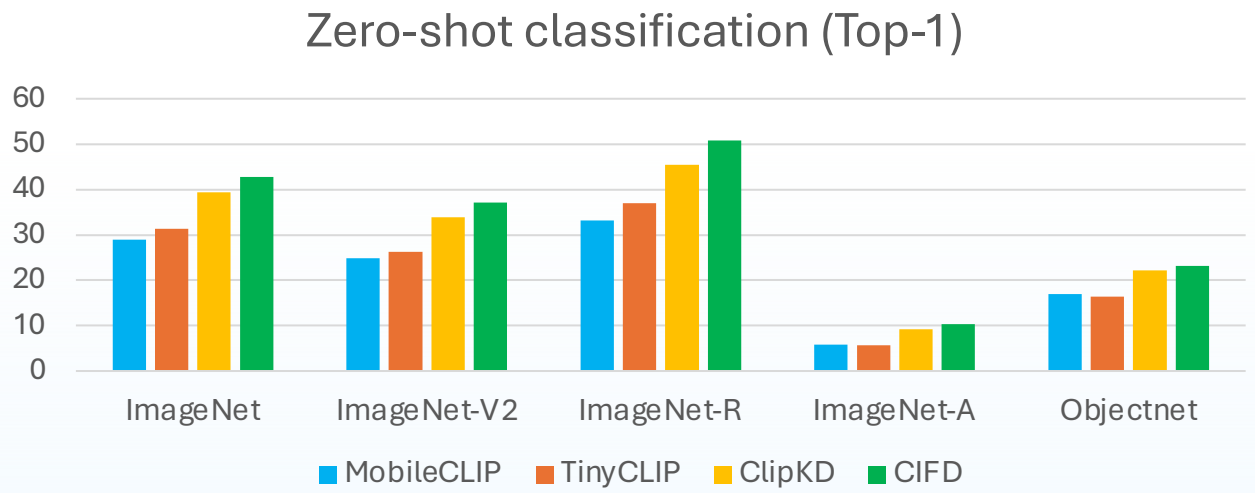


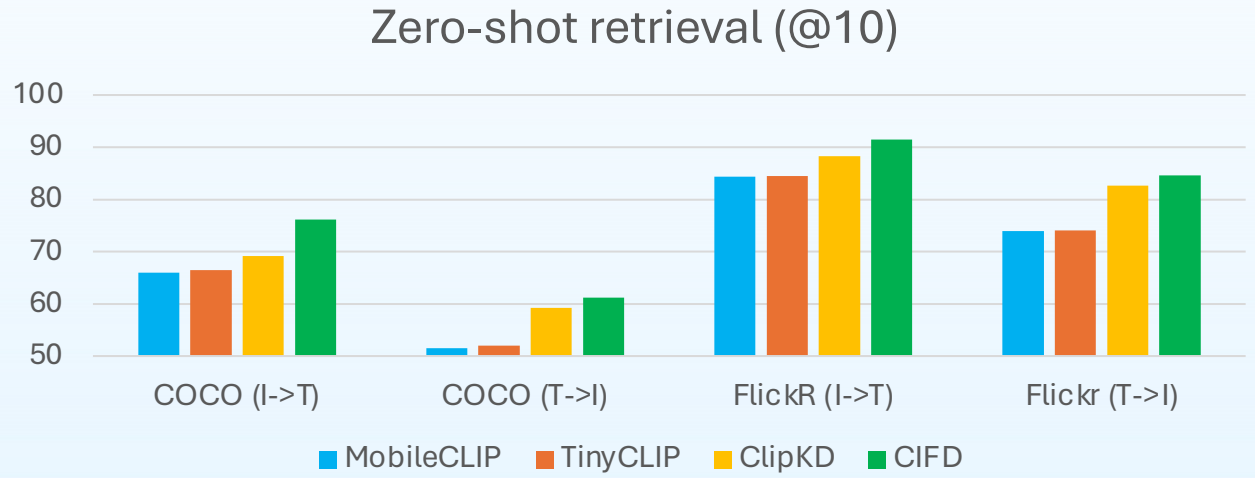


- ➢ CIFD achieves SoTA performance over multiple competing works.
- ➢ CIFD performs well in both cases when teacher and student are from the same arch style or not.

Results on CLIP model distillation

- We compare against CLIP specialized distillation methods.
- Experimental setup (refer paper):
  - Teacher: ViT-L-14
  - Student: ViT-S-16
  - Also tested other students like RN50, ViT-B-16.



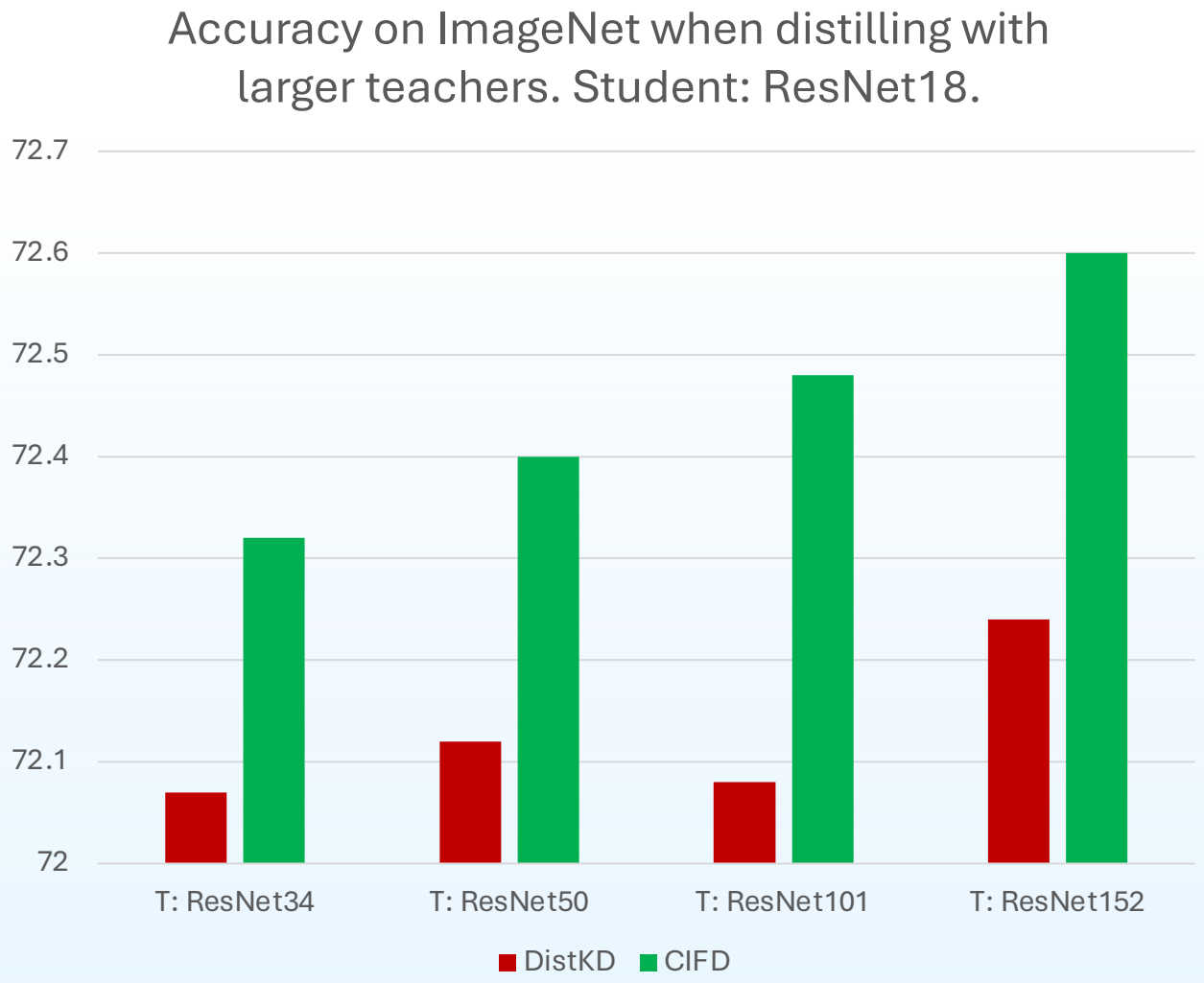


- CIFD beats CLIP specialized distillation methods in CLIP distillation task.
- CIFD achieves best performance across different student-teacher combinations.
- CIFD achieves best performance across zero-shot classification and retrieval.

Results on Large Teacher-Student capacity gap

- Student: ResNet18
- Teacher models and parameter ratio w.r.t. student
  - ResNet34: 1.86
  - ResNet50: 2.19
  - ResNet101: 3.81
  - ResNet152: 5.12

- CIFD shows monotonic performance improvement with teacher size.
- CIFD shows +0.36% improvement over DistKD with RN152 teacher.



Conclusion

- Contributions of the work:
  - Proposed Rate-Distortion-Modules to mimic TAs.
    - RDMs are cheaper to train as they don't relearn low level input features.
    - RDMs can be trained in parallel.
  - Proposed use of IBM for regularizing KD in presence of RDMs.
    - Orthogonally, showed the connection between Information Bottleneck Principle, IBM, and Masked-Image-Modeling.
- Experimental results:
  - Showed superior performance using CIFD for ImageNet classification
  - Showed superior performance distilling CLIP like models over CLIP specific distillation methods.
  - Showed superior performance w.r.t. large student teacher capacity gap.