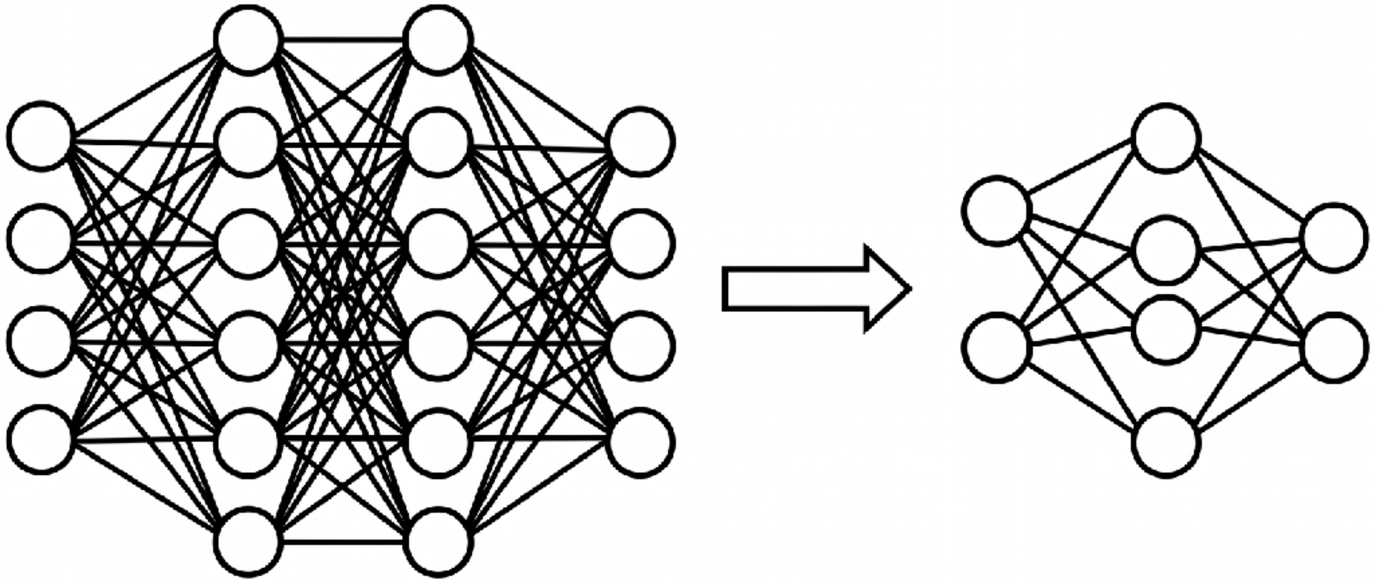
CSCI1470

Deep Learning

Randall Balestriero

Recap

Introduction to Knowledge Distillation Distillation.



Teacher Model
(Large, Accurate)
92%

Student Model
(Small, Efficient)
85%

How can the Student learn from the Teacher's performance?

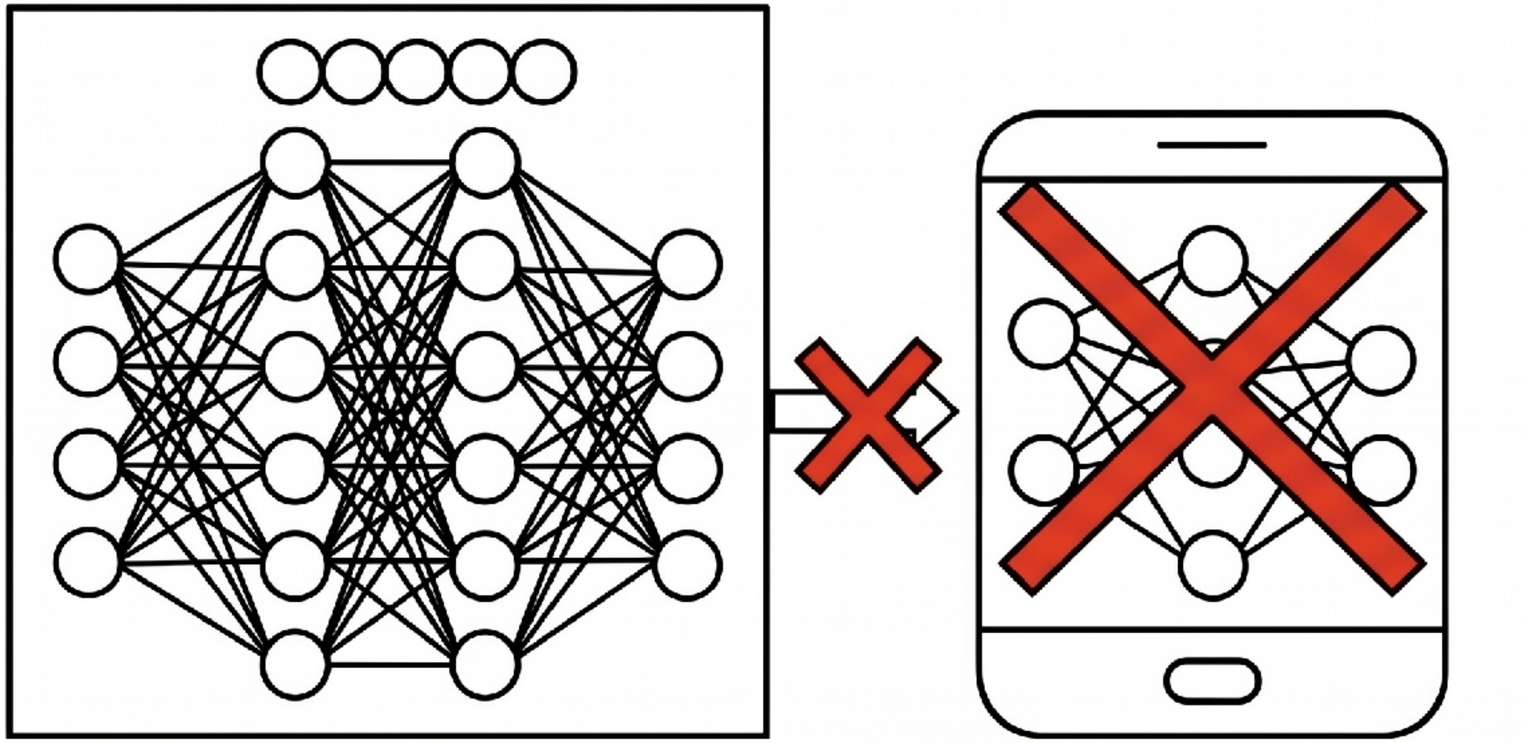
Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

The Core Problem: Why Distill?



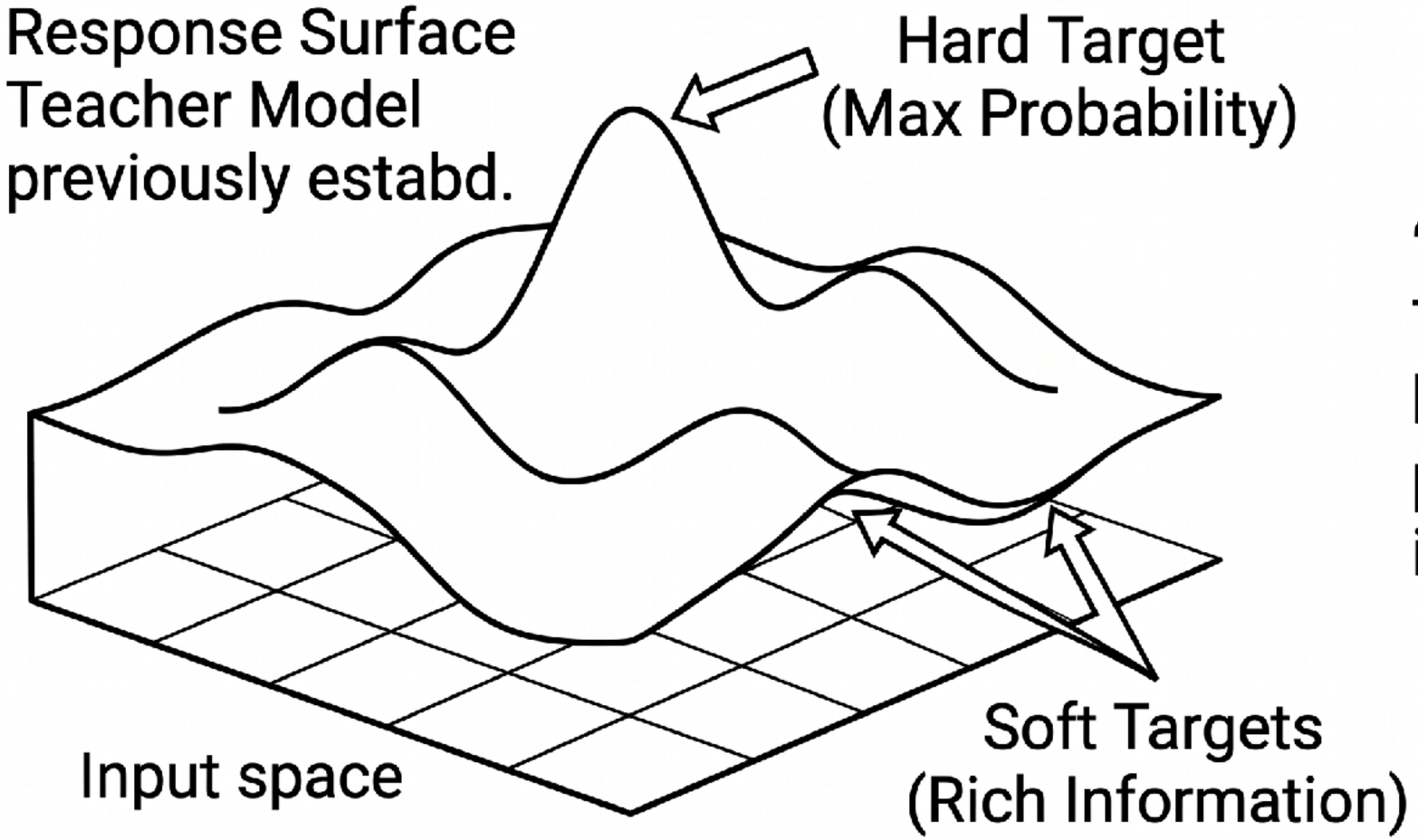
Teacher Ensemble
(Cloud Server)

Mobile Device
(Target Deployment)

- Large models have massive computational costs.
- Deployment on edge devices (mobile, IoT) is restricted by memory and latency.
- Standard training of small models limits accuracy.

⇒ ?

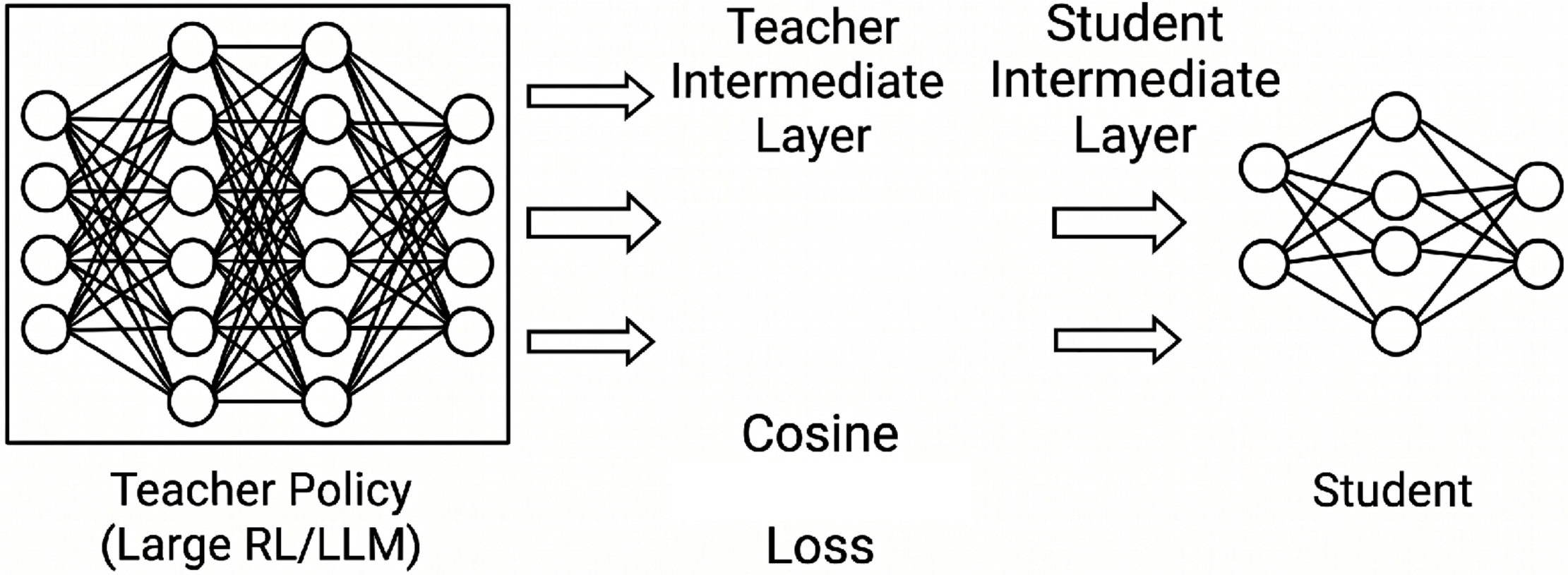
The Response Surface & “Dark Knowledge”



“Dark Knowledge”:
The rich information hidden in the relative probabilities of incorrect classes.

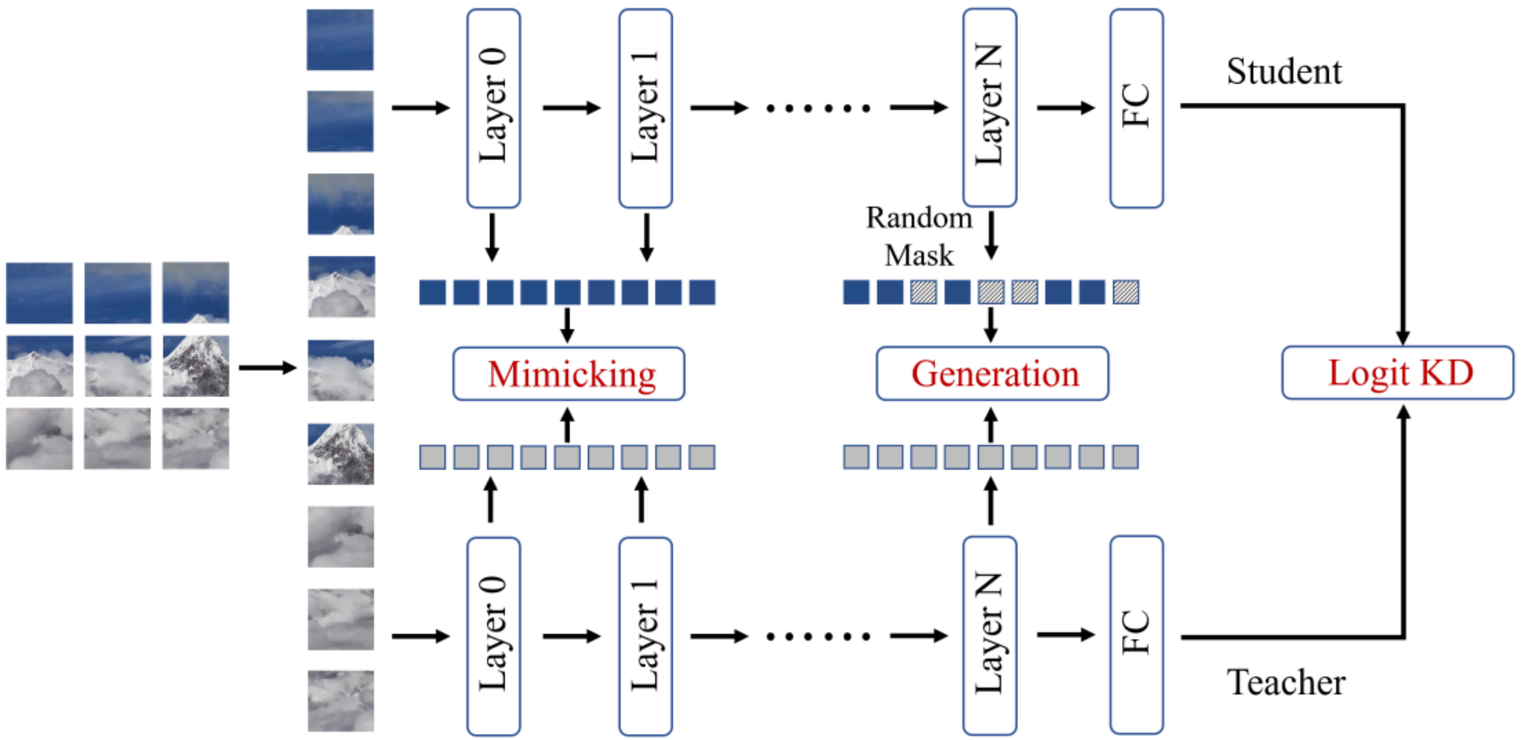
What other “knowledge” can we extract?

Distillation Architecture: Feature Alignment



ViTKD: Feature-based Knowledge Distillation for Vision Transformers

Zhendong Yang[1,2*] Zhe Li[3] Ailing Zeng[2] Zexian Li[4] Chun Yuan[1†] Yu Li[2†]
[1]Tsinghua Shenzhen International Graduate School [2]International Digital Economy Academy (IDEA)
[3]Institute of Automation, Chinese Academy of Sciences [4]Beihang University
yangzd21@mails.tsinghua.edu.cn axel.li@outlook.com lizexian0427@gmail.com
yuanc@sz.tsinghua.edu.cn {zengailing, liyu}@idea.edu.cn



- 10 models pretrained on web data (LVD-1689M dataset)
  - 1 ViT-7B trained from scratch,
  - 5 ViT-S/S+/B/L/H+ models distilled from the ViT-7B,
  - 4 ConvNeXt-{T/S/B/L} models distilled from the ViT-7B,
- 2 models pretrained on satellite data (SAT-493M dataset)
  - 1 ViT-7B trained from scratch
  - 1 ViT-L distilled from the ViT-7B

Improving Group Fairness in Knowledge Distillation via Laplace Approximation of Early Exits

Edvin Fasth
EESC
KTH Royal Institute of Technology
edvinfa@kth.se

Sagar Singh
CSE
IIT Bombay
sagarsingh@cse.iitb.ac.in

Fairness Implications of GNN-to-MLP Knowledge Distillation

Margaret Capetz
Department of Computer Science
UCLA
mcapetz17@g.ucla.edu

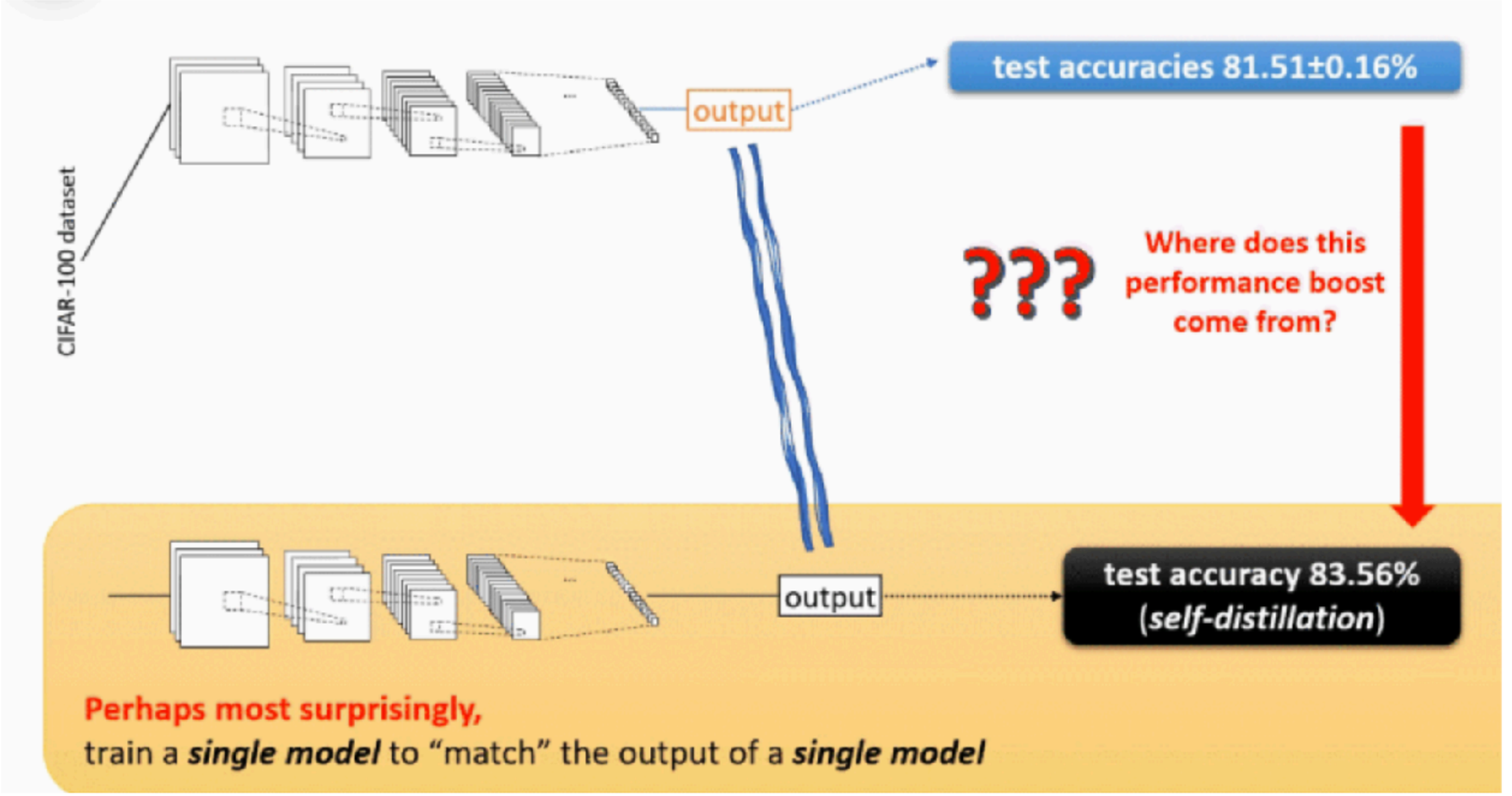
Yizhou Sun
Department of Computer Science
UCLA
yzsun@cs.ucla.edu

Arjun Subramonian
Department of Computer Science
UCLA
arjunsub@g.ucla.edu

TOWARDS UNDERSTANDING ENSEMBLE, KNOWLEDGE DISTILLATION AND SELF-DISTILLATION IN DEEP LEARNING

Zeyuan Allen-Zhu
Meta FAIR Labs
zeyuanallenzhu@meta.com

Yuanzhi Li
Mohamed bin Zayed University of AI
Yuanzhi.Li@mbzuai.ac.ae



Analyzing class confusion types

Fine-grained and non-trivial

wheel

car

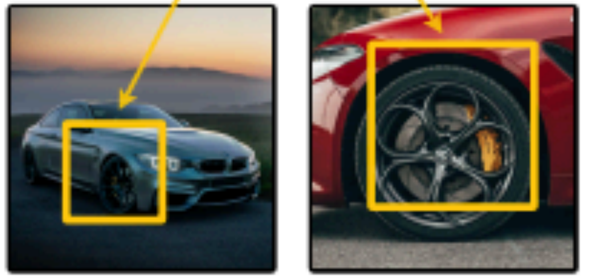
Ambiguous

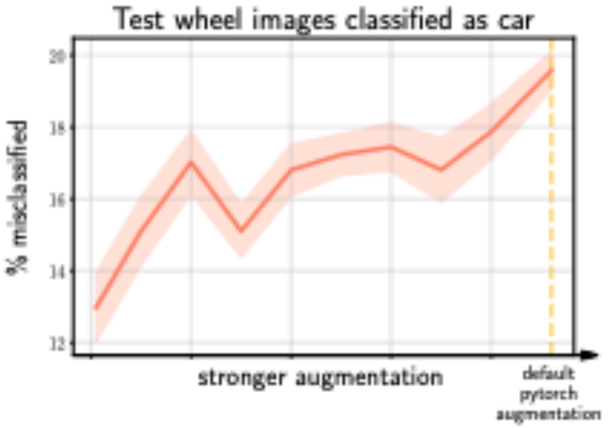
sunglass

sunglasses

Accuracy degradation is caused by class confusions induced by augmentation

Default pytorch Random Resized Crop
(input to the model)



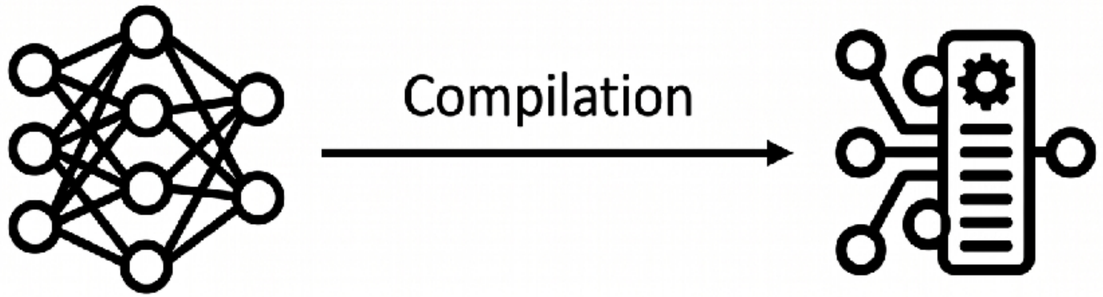


Augmentation strategy which accounts for class interactions leads **+2.5%** improvement on affected ImageNet classes

Introduction to Model Compilation

Trained Model
(Framework: PyTorch/TF)

Compiled Engine
(Hardware Targeted)



Trained Model
(Framework: PyTorch/TF)

Compiled Engine
(Hardware Targeted)

- Definition: Transforming a high-level model definition into into a specialized execution engine.
- Objective: Optimize for target hardware (CPU, GPU, Edge).
- Result: Lower latency and higher throughput during inference.

A few examples of emerging edge AI applications

Edge AI use case

In-home smart cameras can recognize that a person(s) has entered an area

nest

aws

On-device facial recognition and object recognition, where user data doesn`t leave the device

HUAWEI

On-board AI making instantaneous driving decisions

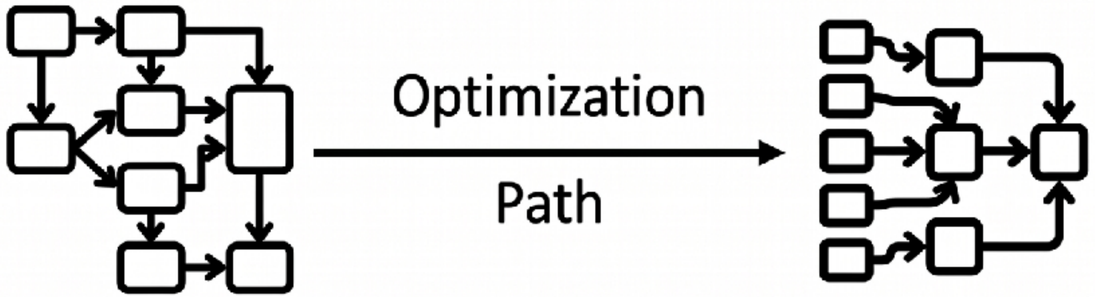
Vision for baby monitors, drones, robots, and other devices that can respond to situations without internet connection

intel

Cloud stores large datasets, trains algorithms, collects edge data, pushes AI model updates

Common Compilation Frameworks

ONNX (Open Neural Network Exchange)



TensorRT (NVIDIA High-Perf. Inference)

Interoperable format, framework agnostic.

NVIDIA GPU-specific, aggressive optimization.

- Definition: Transforming a high-level model definition into into a specialized execution engine.
- Objective: Optimize for target hardware (CPU, GPU, Edge).
- Result: Lower latency and higher throughput during inference.

Easy, fast, and cheap LLM serving for everyone

Thank you!