Knowledge Distillation for Building Lightweight Deep Learning Models in Visual Classification Tasks

ECE 1512: DIGITAL IMAGE PROCESSING AND APPLICATIONS

SEMESTER: WINTER 2022 PROJECT “B” TUTORIAL

PRESENTER: AHMAD SAJEDI

Outline

A. Tutorial on Knowledge Distillation (KD)
   - Motivation
   - Approaches for Knowledge Distillation Framework

B. Project “B” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

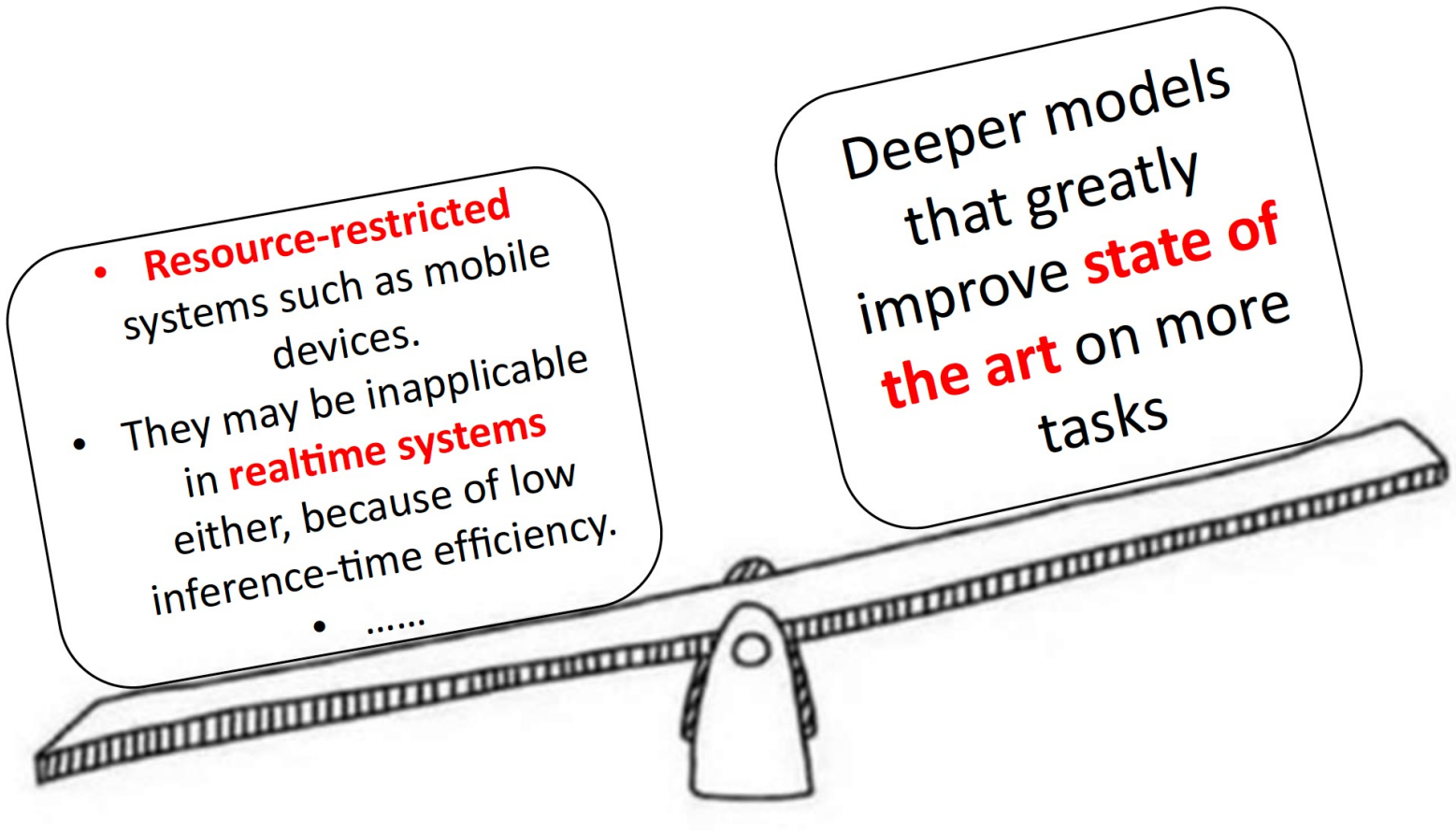
Outline

A. Tutorial on Knowledge Distillation (KD)
   - Motivation
   - Approaches for knowledge distillation framework

B. Project “B” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

Motivation


• Resource-restricted systems such as mobile devices.
• They may be inapplicable in realtime systems either, because of low inference-time efficiency.
• .....
Deeper models that greatly improve state of the art on more tasks

Motivation

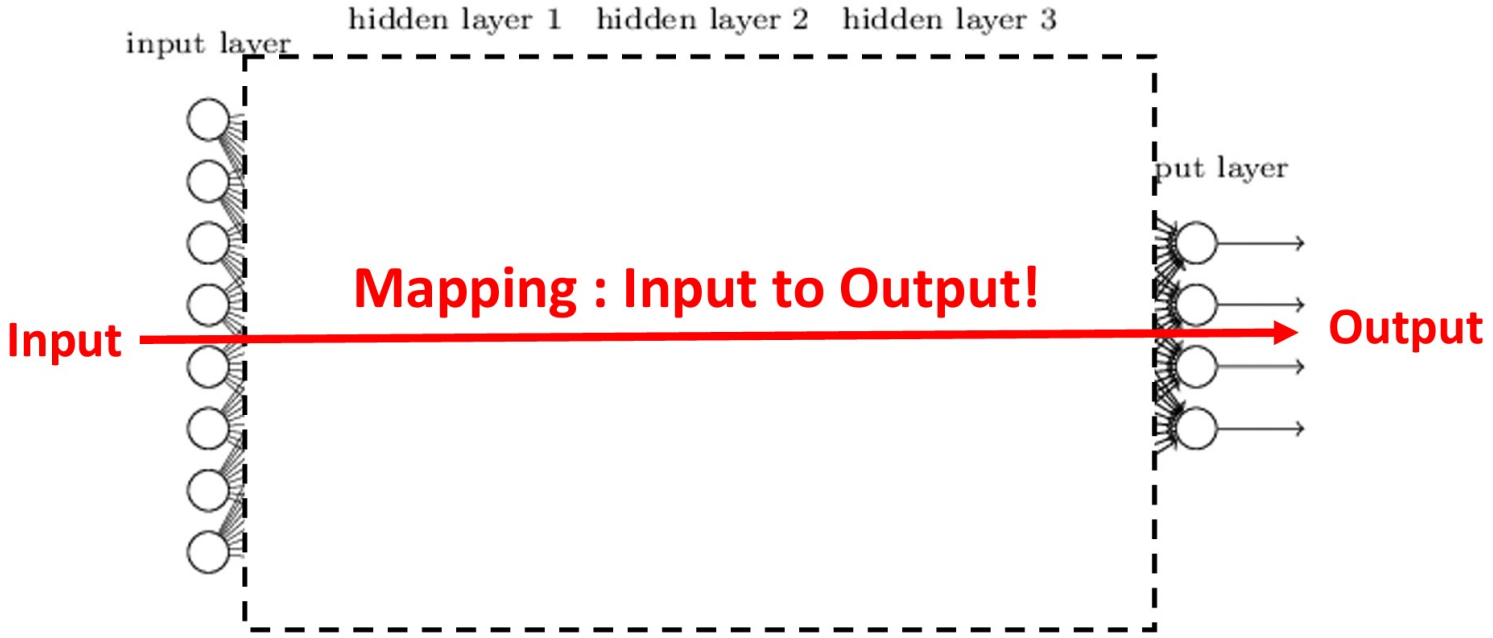
Model Compression:

- Goal: make a lightweight model that is fast, memory-efficient, and energy-efficient
- Especially useful for **edge device** such as mobile device.

Several flavor:

- Whether training a lightweight model or compressing a trained model
- Different techniques:
    1. Sparse Regularization
    2. Quantization
    3. Weight Sharing
    4. Pruning
    5. **Knowledge Distillation**

What is Knowledge Distillation?



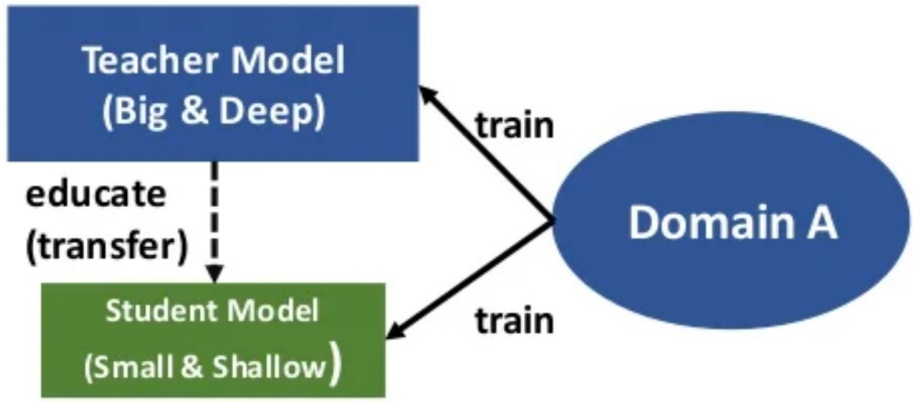
A more abstract view of the knowledge, that frees it from any instantiation, is that it is **a learned mapping from input vectors to output vectors.**

What is Knowledge Distillation?

Knowledge distillation is a process of distilling or transferring the knowledge from a (set of) large, cumbersome model(s) to a lighter, easier-to-deploy single model, without significant loss in performance.

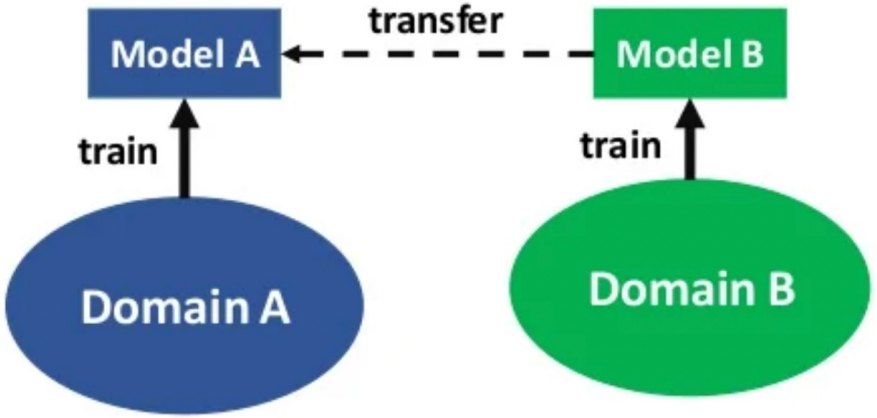
Knowledge Distillation vs. Transfer Learning

Knowledge Distillation (Transfer)



- For model compression
- To improve performance of student over teacher

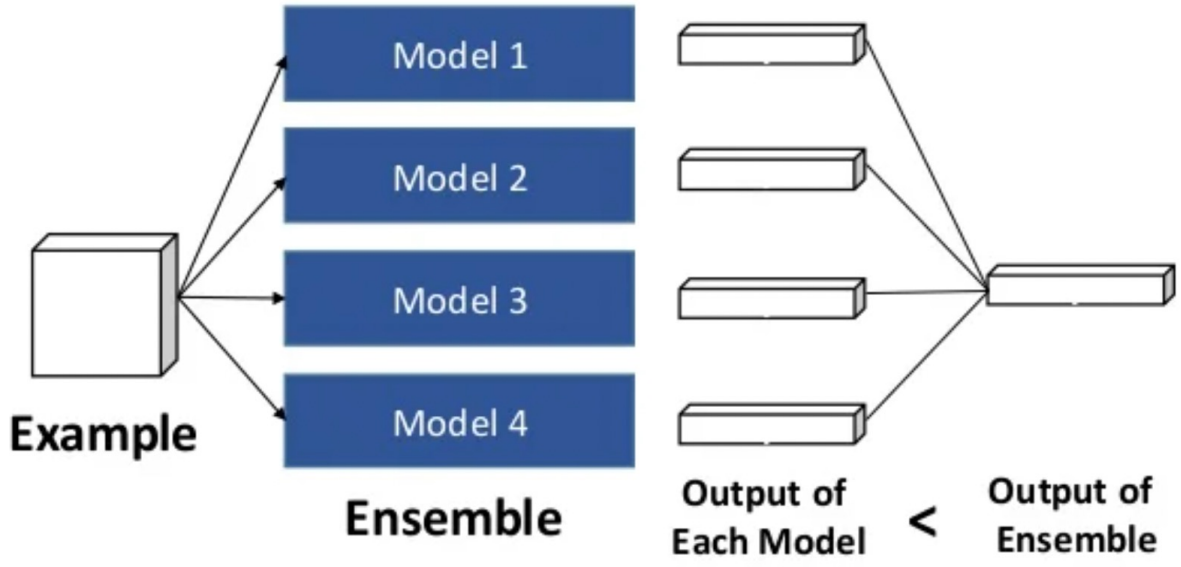
Transfer Learning



- When data is not sufficient.
- When label for a problem is not presented.
- E.g., pretrained-model on ImageNet

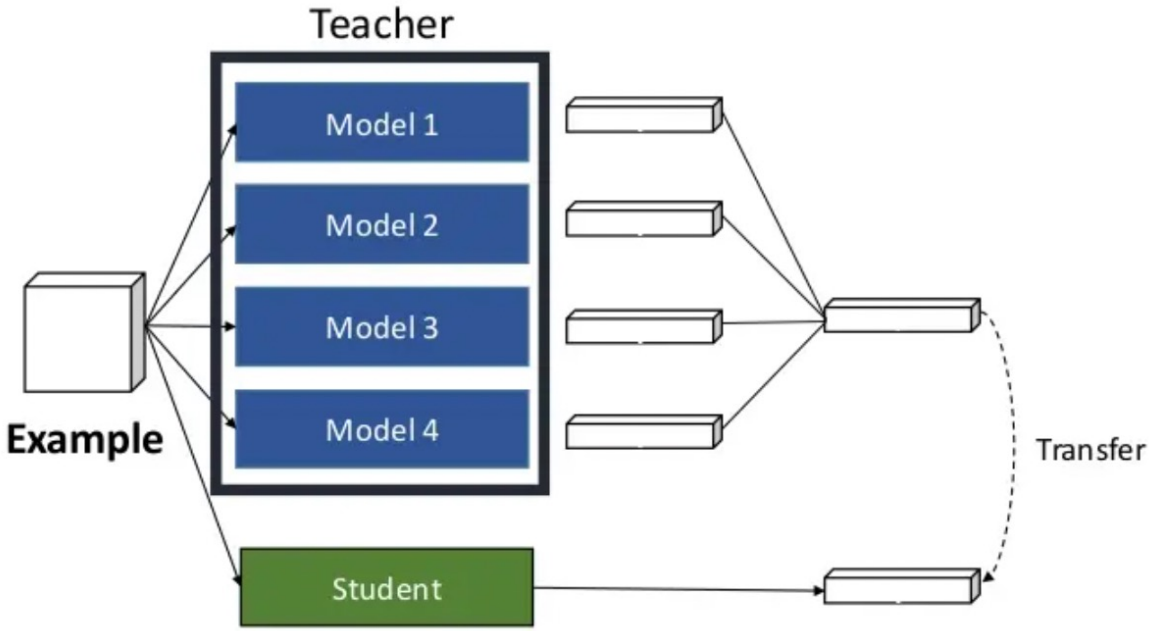
Model Compression Using Knowledge Distillation

- Ensemble is an easy way to improve performance of a Neural Network.
- However, it requires large computing resources.

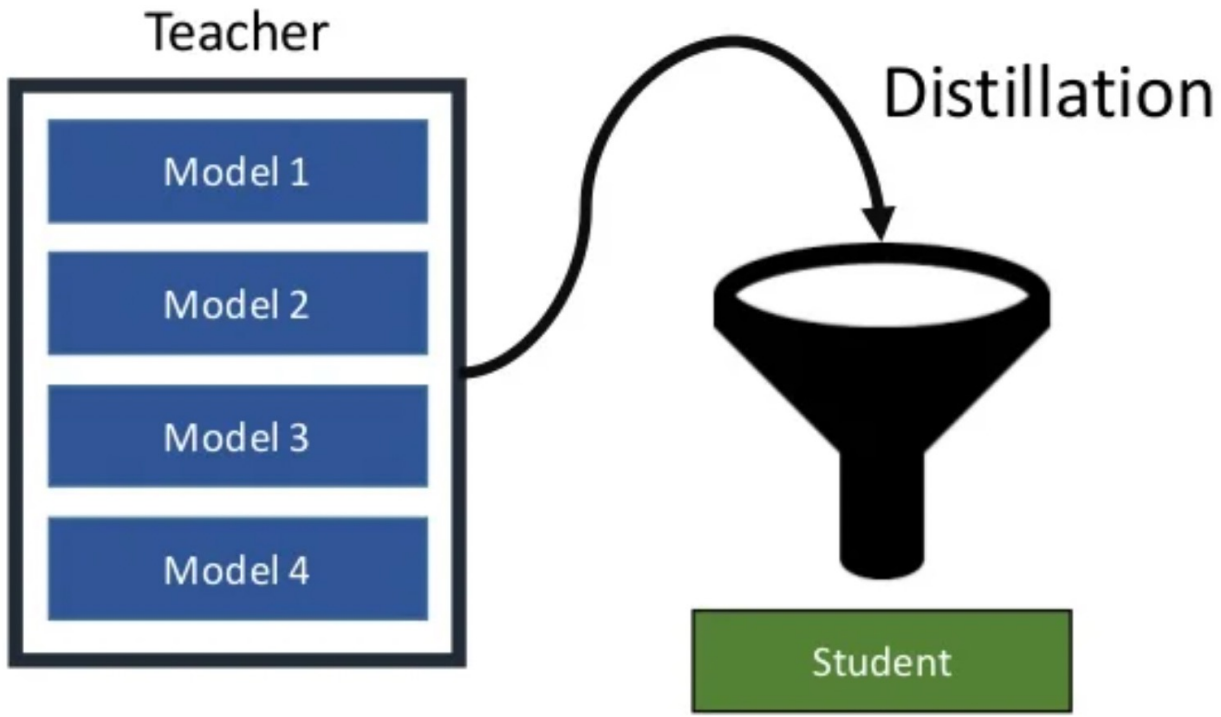


Model Compression Using Knowledge Distillation

- By educating the student model to mimic output of the teacher model, the student model can achieve comparable performance.



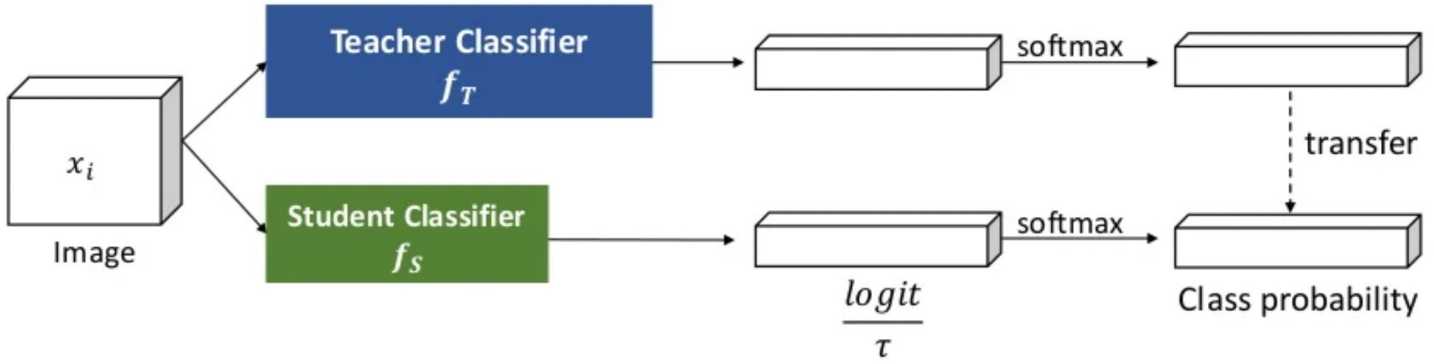
Model Compression Using Knowledge Distillation



Recent Approaches: Transfer Class Probability

- **Distilling the knowledge in a Neural Network**
  Hinton *et al.* In *NIPS*, 2014

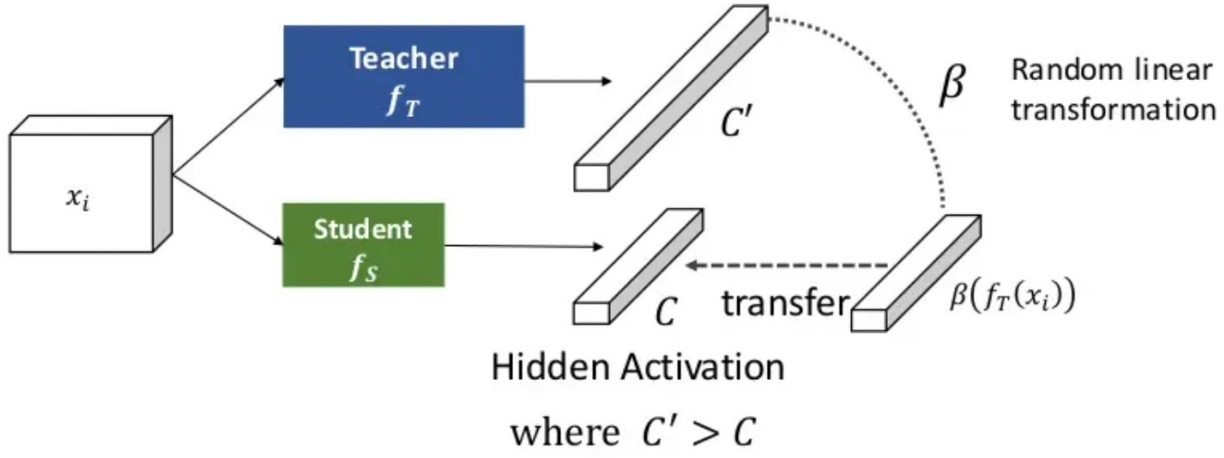
Distillation Objective: $\sum_{x_i \in \mathcal{X}} KL(softmax\left(\frac{f_T(x_i)}{\tau}\right), softmax\left(\frac{f_S(x_i)}{\tau}\right))$



Recent Approaches: Transfer Hidden Activation

- **FitNets: Hints for Thin Deep Nets**
  Romero *et al.* In *ICLR,* 2015

Distillation Objective: $\sum_{x_i \in \mathcal{X}} \| \beta f_T(x_i) - f_S(x_i) \|_2^2$

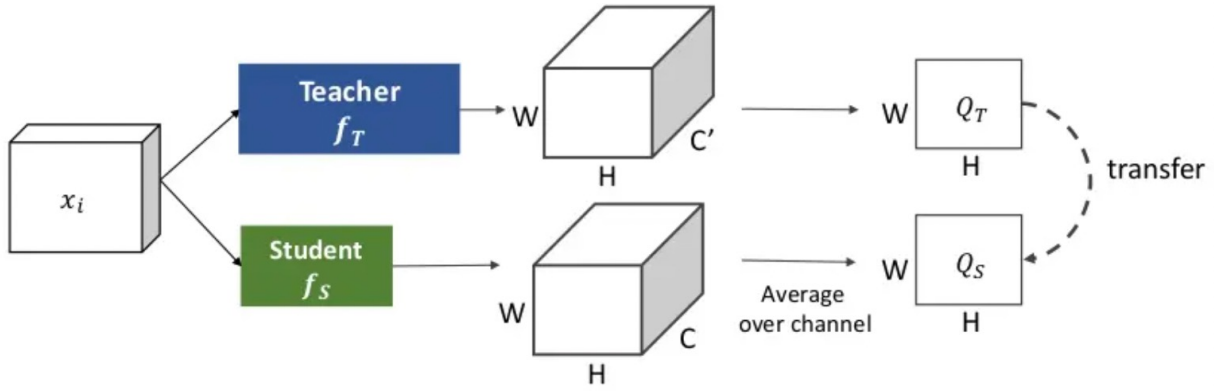


Recent Approaches: Transfer Attention

- **Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer**

  Zagoruyko & Komodakis. In *ICLR,* 2017

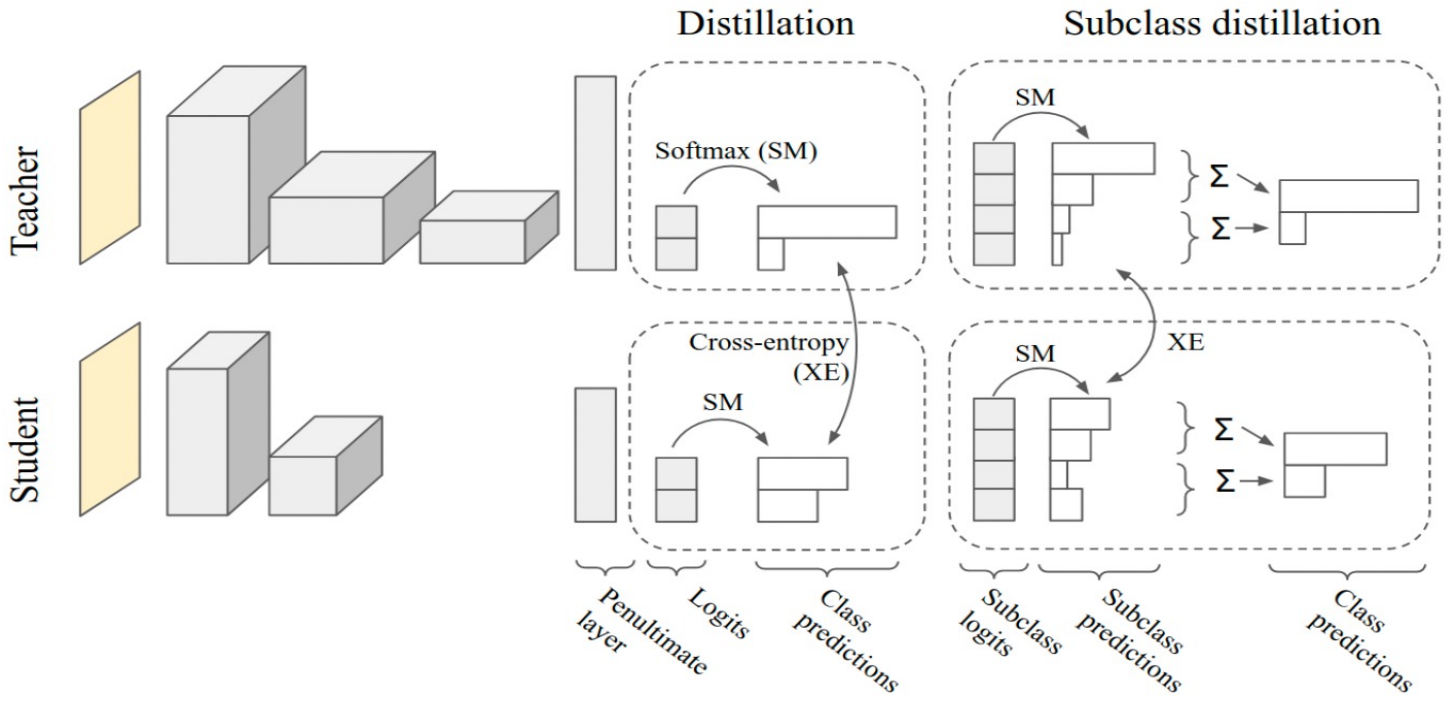
Distillation Objective: $\sum_{x_i \in \mathcal{X}} \left\| \frac{Q_T^i}{\|Q_T^i\|} - \frac{Q_S^i}{\|Q_S^i\|} \right\|_2$



Recent Approaches: Transfer Subclass Knowledge

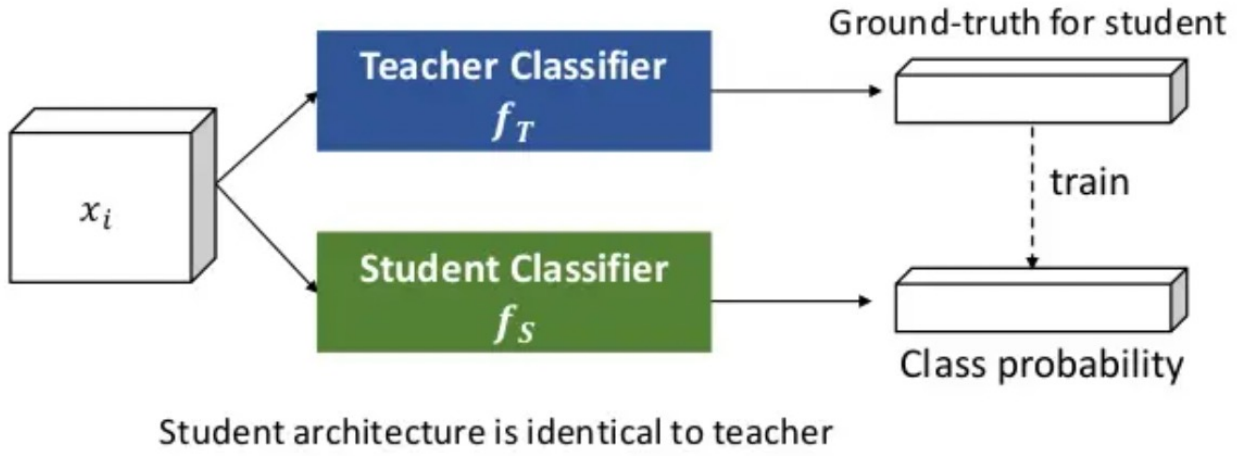
- **Subclass Distillation**

  Muller et al. In *arXiv,* 2020.



Recent Approaches: Student Over Teacher

- **Born-Again Neural Networks** (Furlanello *et al.* In *ICML*, 2018.)
- **Label Refinery: Improving ImageNet Classification through Label Progression** (Bagherinezhad *et al.* In *arXiv*, 2018.)



Surprisingly, the student is significantly better than the teacher.

Outline

A. Tutorial on Knowledge Distillation (KD)
   - Motivation
   - Approaches for knowledge distillation framework

B. Project “B” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

Project Goal

- Inspect the ability of knowledge distillation methods in model compression for CNNs in two different scenarios.
- **Specifications:**
  - Datasets:
    - MNIST: Generic digit pattern classification
    - MHIST: Histopathological tissue classification
  - Models:
    - MNIST dataset: Teacher = CNN with 2 conv. layers, Student = Fully connected.
    - MHIST dataset: Teacher = ResNet50V2, Student = MobileNetV2
  - Evaluation metrics:
    - Test Accuracy%
    - F1-Score, AUC%
    - FLOPs

Dataset: MNIST

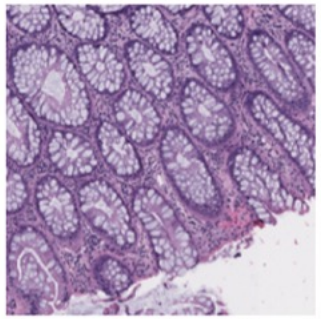
- Multi-class digit classification dataset
- The MNIST dataset is divided into 10 classes, each of which represents a digit between 0-9.
- The digits have been size-normalized and centered in a fixed-size image.
- 60000 train data +10000 test data.

http://yann.lecun.com/exdb/mnist/

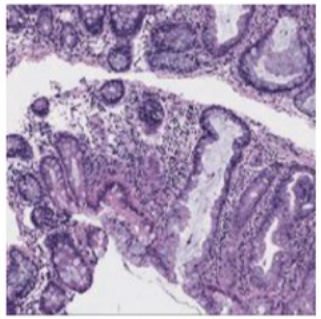
Dataset: Minimalistic HIStopathology (MHIST)

- Binary-class texture analysis in colorectal cancer histology.
- classes:
  - (a) Hyperplastic Polyp (benign),
  - (b) Sessile Serrated Adenoma (precancerous).
- 2175 train data + 977 test data.
- Not equally-balanced dataset:
  - 2162 images per class HP
  - 990 images per class SSA

Binary classification task



Hyperplastic Polyp (HP)
- Benign



Sessile Serrated Adenoma (SSA)
- Precancerous

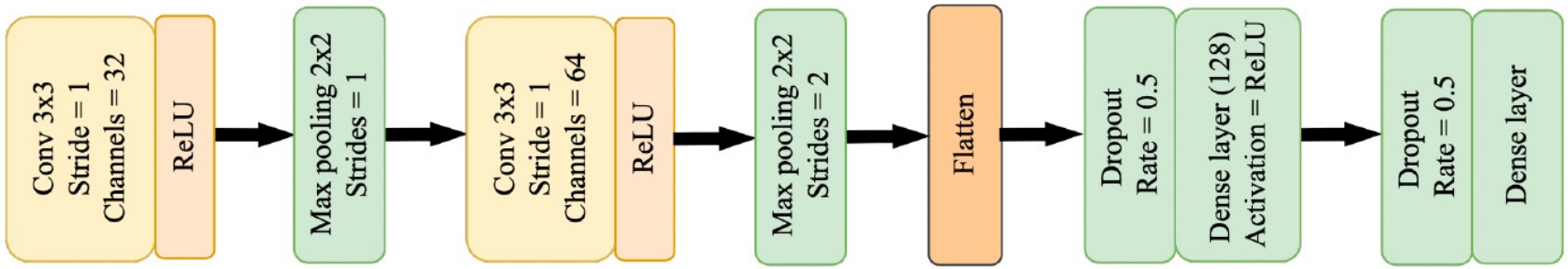
Dataset summary

Dataset size	*N = 3,152*
Image size	224 x 224 pixels
Disk space	354 MB
Ground-truth labels	Majority vote of seven pathologists

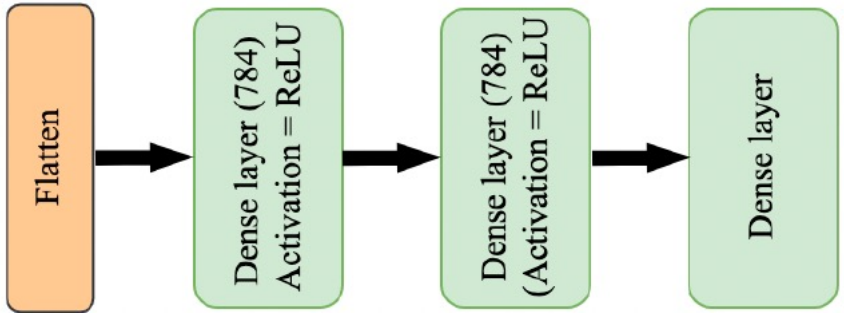
https://arxiv.org/abs/2101.12355

Models for MNIST dataset

- Teacher Model:

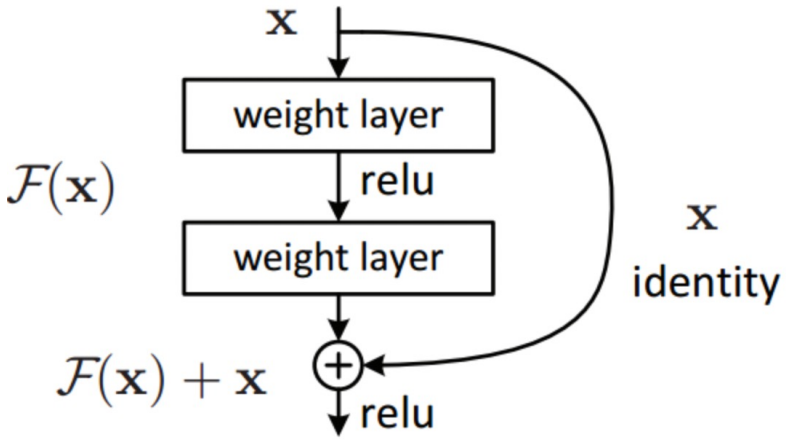


- Student Model:



Models for MHIST dataset

- Teacher Model: **Pre-trained ResNet50V2**



Residual Block

- Student Model: **Pre-trained MobileNetV2**
  - MobileNetV2 is **a convolutional neural network architecture that seeks to perform well on mobile devices.**

Note: In this task, you should use **transfer learning** for training the models.

Evaluation Metrics

- **Model Performance:**
  - MNIST Dataset: Test Accuracy
  - MHIST Dataset: F1-Score, AUC (Evaluation metrics should be suitable for imbalanced dataset)
- **Model Complexity:**
  - Floating point operations (FLOPs): is the number of floating point operations, it means the amount of calculation, it can be used to measure the **algorithm/ Model complexity**.
  - A floating point operation is **any mathematical operation (such as +, -, *, /) or assignment that involves floating-point numbers.**

Outline

A. Tutorial on visual explainable AI (XAI)
   - Motivation
   - Primer on explainability in Artificial Intelligence (AI)
   - Approaches for visual explanation generation

B. Project “A” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

THANK YOU

Questions?