Knowledge Distillation

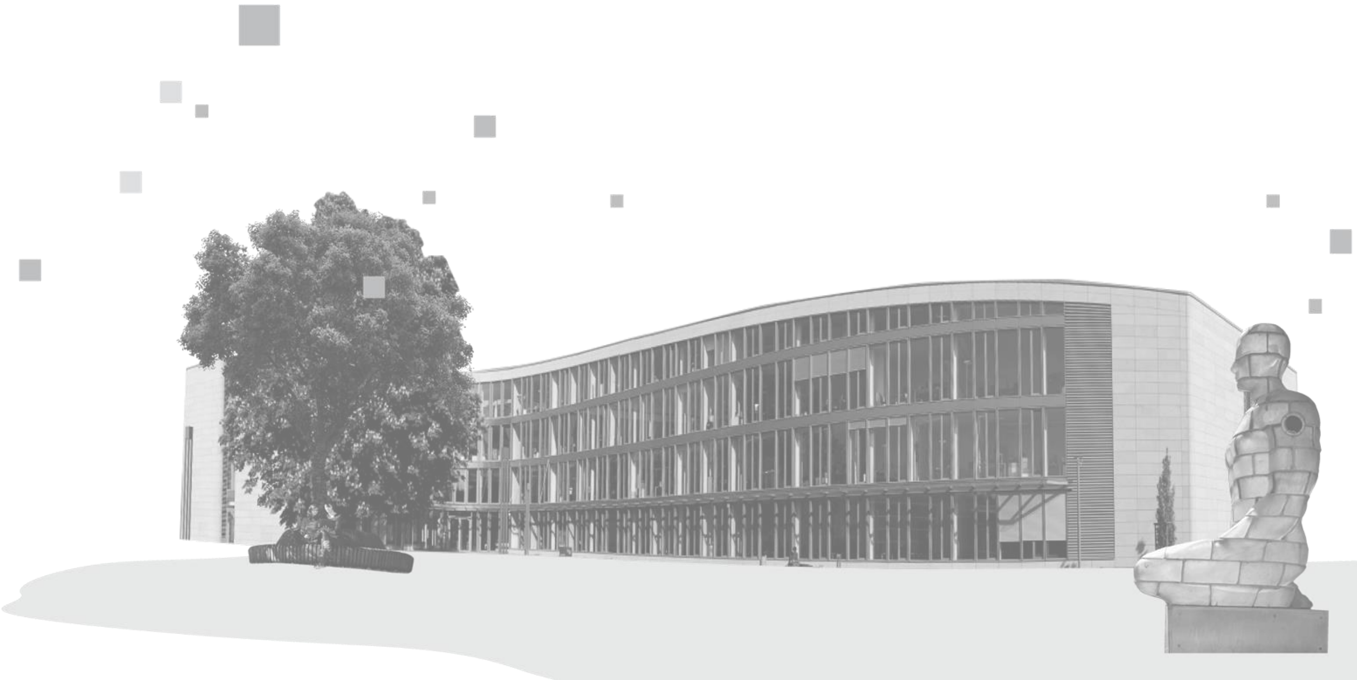
PD Dr. Haojin Yang

Multimedia and Maschine Learning Group

Hasso Plattner Institute

Design IT.
Create Knowledge.

www.hpi.de



What is Knowledge Distillation?

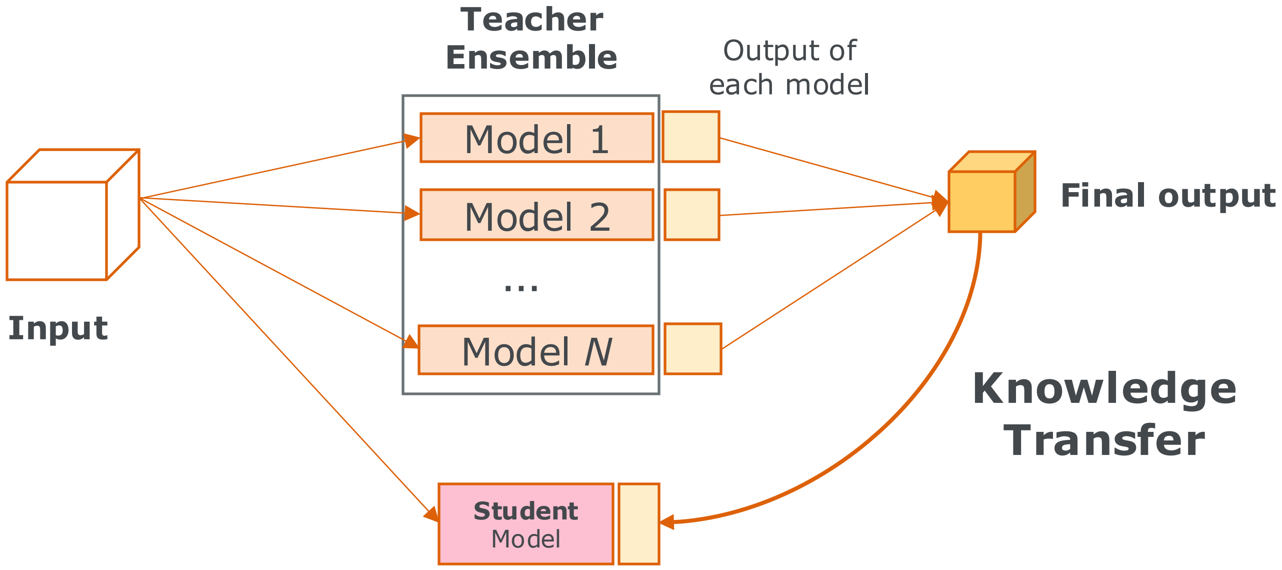
- Many insects have a larval and adult form.

- **Training** and **deployment** phases of large-scale machine learning systems
  - Different requirements on latency and computational resources
  - Therefore, the model can also have different forms at different phases
  - A method of transformation between forms is called "Knowledge Distillation"

Hinton, Geoffrey et al. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

What is Knowledge Distillation?

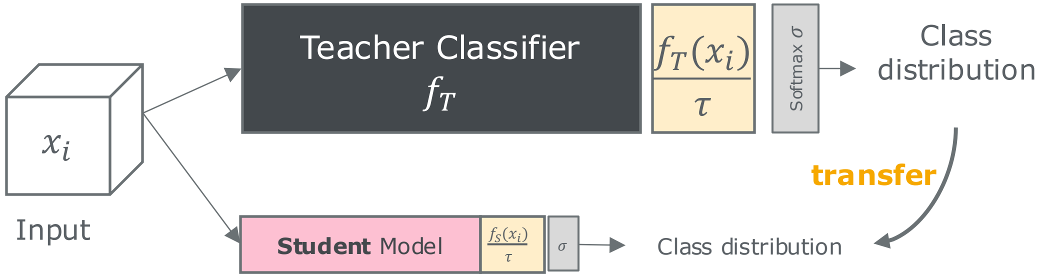
- A model ensemble improves performance but requires large computing resources.
- Buciluǎ et al. demonstrate that the knowledge acquired by a large ensemble of models can be transferred to a single small model.



Buciluǎ et al. "Model compression." ACM KDD 2006

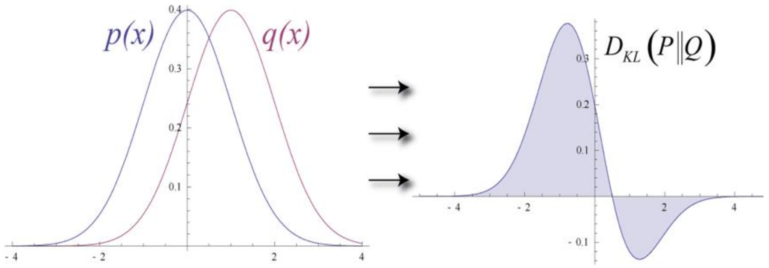
What Knowledge can be Transferred?

- Hinton et al. consider the neural network model a black box; knowledge can be regarded as the mapping from input to output.
- Transferring knowledge from a big model to a small model → **Model Compression**



Kullback–Leibler divergence as distillation loss:

$$D_{KL} = \sum_{x_i \in X} KL(softmax\left(\frac{f_T(x_i)}{\tau}\right), softmax\left(\frac{f_S(x_i)}{\tau}\right))$$



Mundhenk at English Wikipedia, CC BY-SA 3.0, via Wikimedia Commons

Hinton, Geoffrey et al. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

Why Distribution is Better than Hard Label?

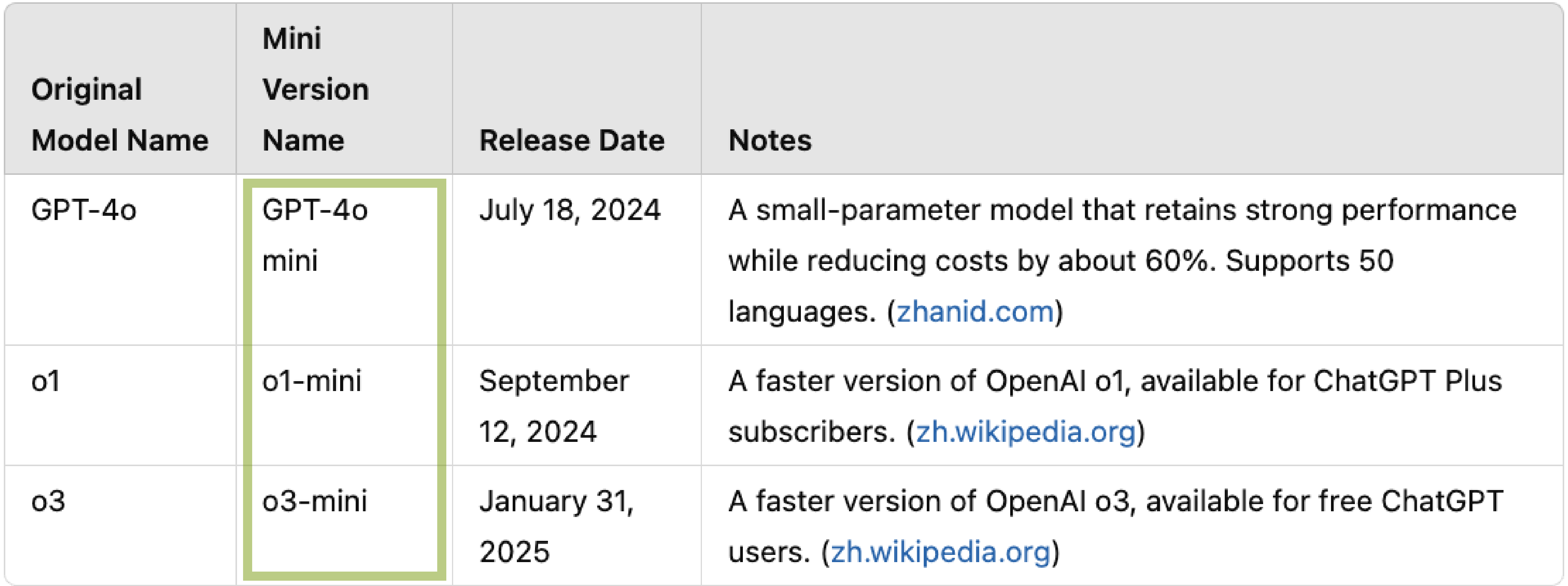
- The class distribution or “soft labels” can offer more information about data than hard labels.
- Soft labels have less gradient variance → smoother training, faster convergence
- Reduces the reliance on labeled data

Class	ID	Hard label
eagle	0	[1,0,0,0]
sparrow	1	[0,1,0,0]
airplane	2	[0,0,1,0]
car	3	[0,0,0,1]

Soft label

[0.65, 0.25, 0.10, 0]

Knowledge Distillation for Large Language Model



Original Model Name	Mini Version Name	Release Date	Notes
GPT-4o	GPT-4o mini	July 18, 2024	A small-parameter model that retains strong performance while reducing costs by about 60%. Supports 50 languages. (zhanid.com)
o1	o1-mini	September 12, 2024	A faster version of OpenAI o1, available for ChatGPT Plus subscribers. (zh.wikipedia.org)
o3	o3-mini	January 31, 2025	A faster version of OpenAI o3, available for free ChatGPT users. (zh.wikipedia.org)

Summary

- KD is a method for model compression
  - Utilizes the fact that large models are often overparameterized
- Teacher-student paradigm
  - Student model learns based on soft labels produced by its teacher
  - Soft labels have multiple benefits
- KD can reduce energy consumption of AI computing
  - Train faster
  - Student model for deployment