Distilling the Knowledge in a Neural Network

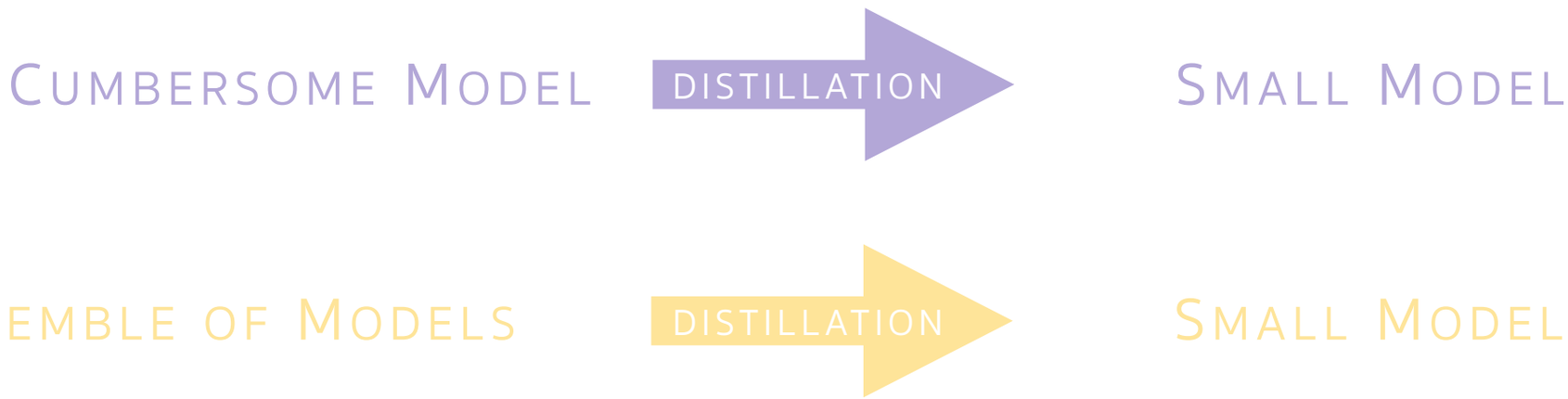
Katya

7 April 2022

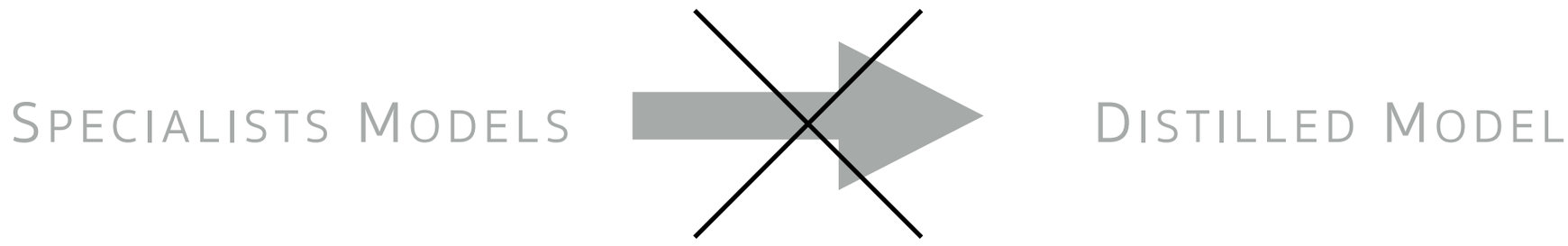
Paper Overview

arXiv1503.02531

Topics covered in the paper



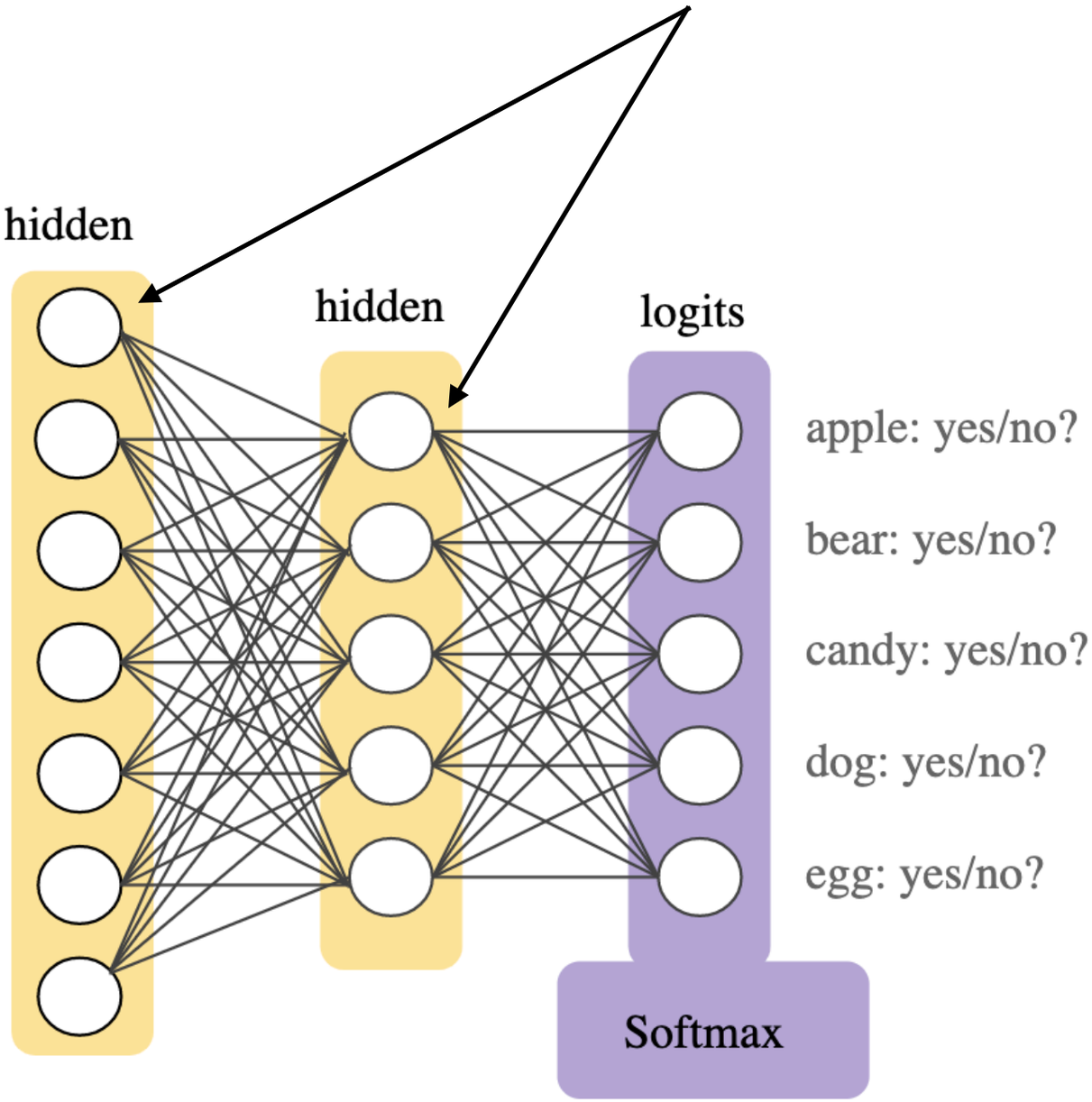
They showed that it is also possible to distill the knowledge of ensemble of models into a single model that works significantly better than a model of the same size that is learned directly from the same training data



The model from Google takes ~ 6 month to train, so they come up with an idea to train copies of the main classification model on special classes in the dataset, so it becomes a specialist in the particular type of inputs

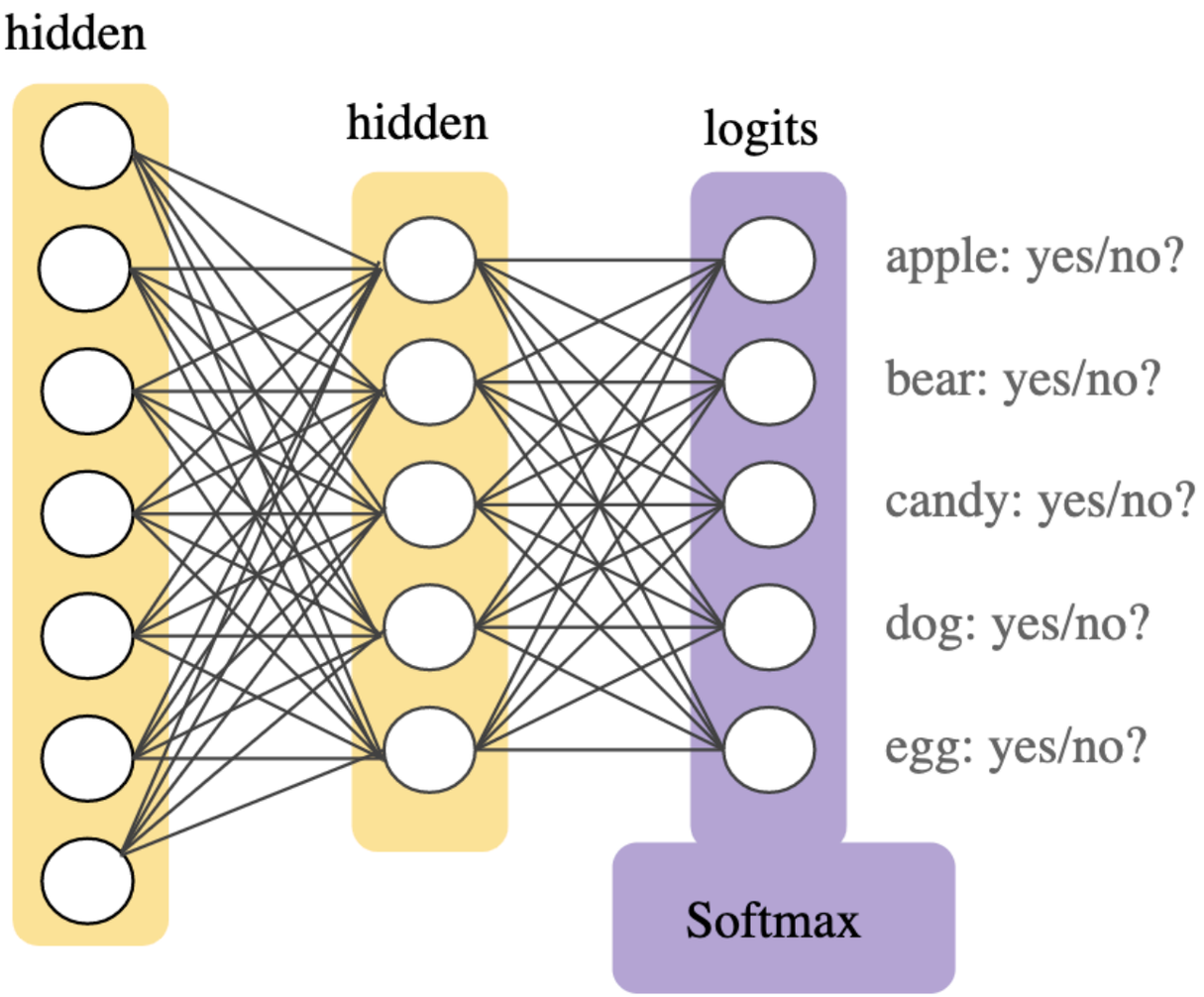
ARXIV1503.02531

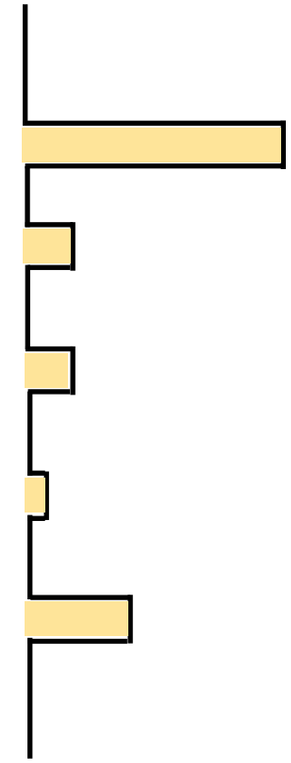
Conceptual block It is usually thought that **knowledge** — learned network parameters



ARXIV1503.02531

A more abstract view of the **knowledge** — learned mapping from input vectors to output vectors





Knowledge Distillation (KD)

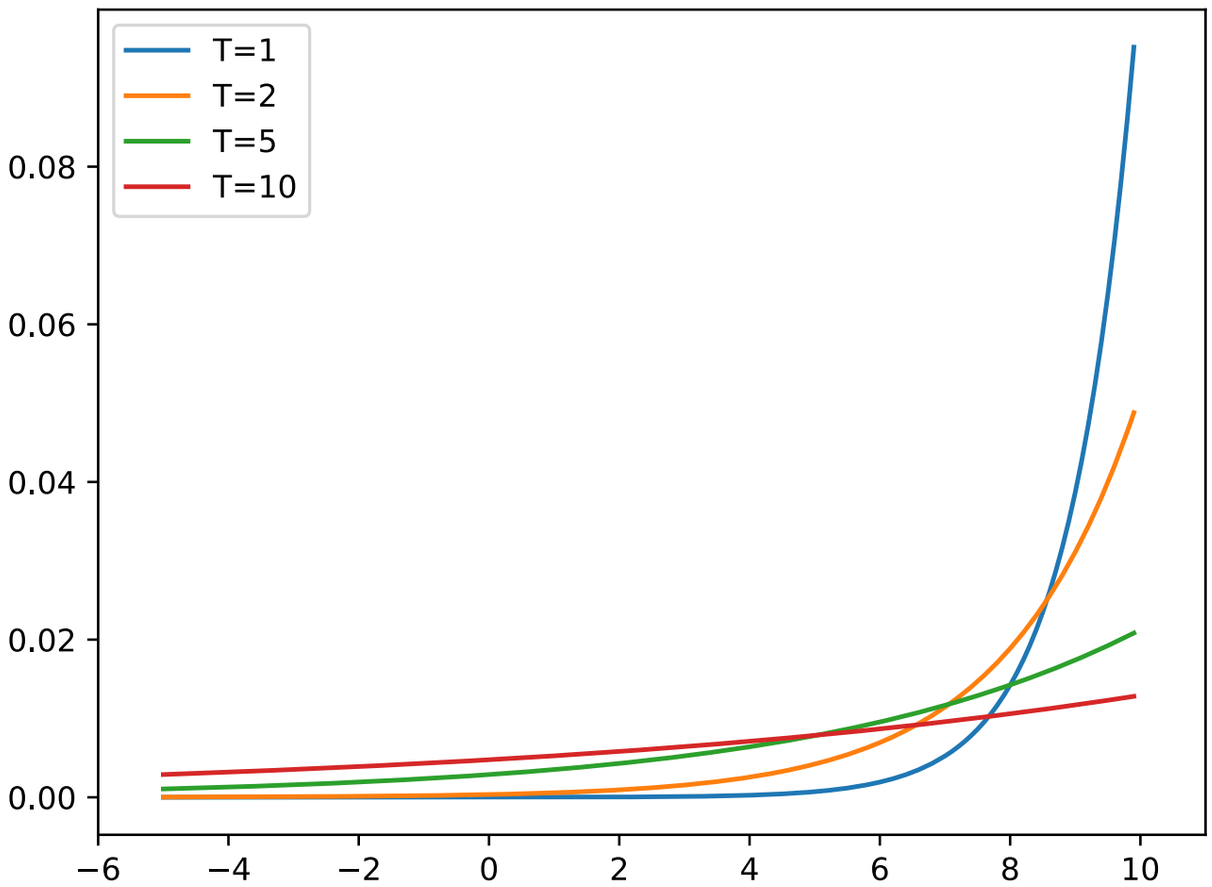
arXiv1503.02531

Keras example

Distillation - raising `Softmax()` temperature **T**, so that the output is suitable soft set of targets

Matching **logits** is a special case of distillation (in the high temperature limit)

`Softmax` = $\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$, where T - temperature parameter



Knowledge Distillation (KD)

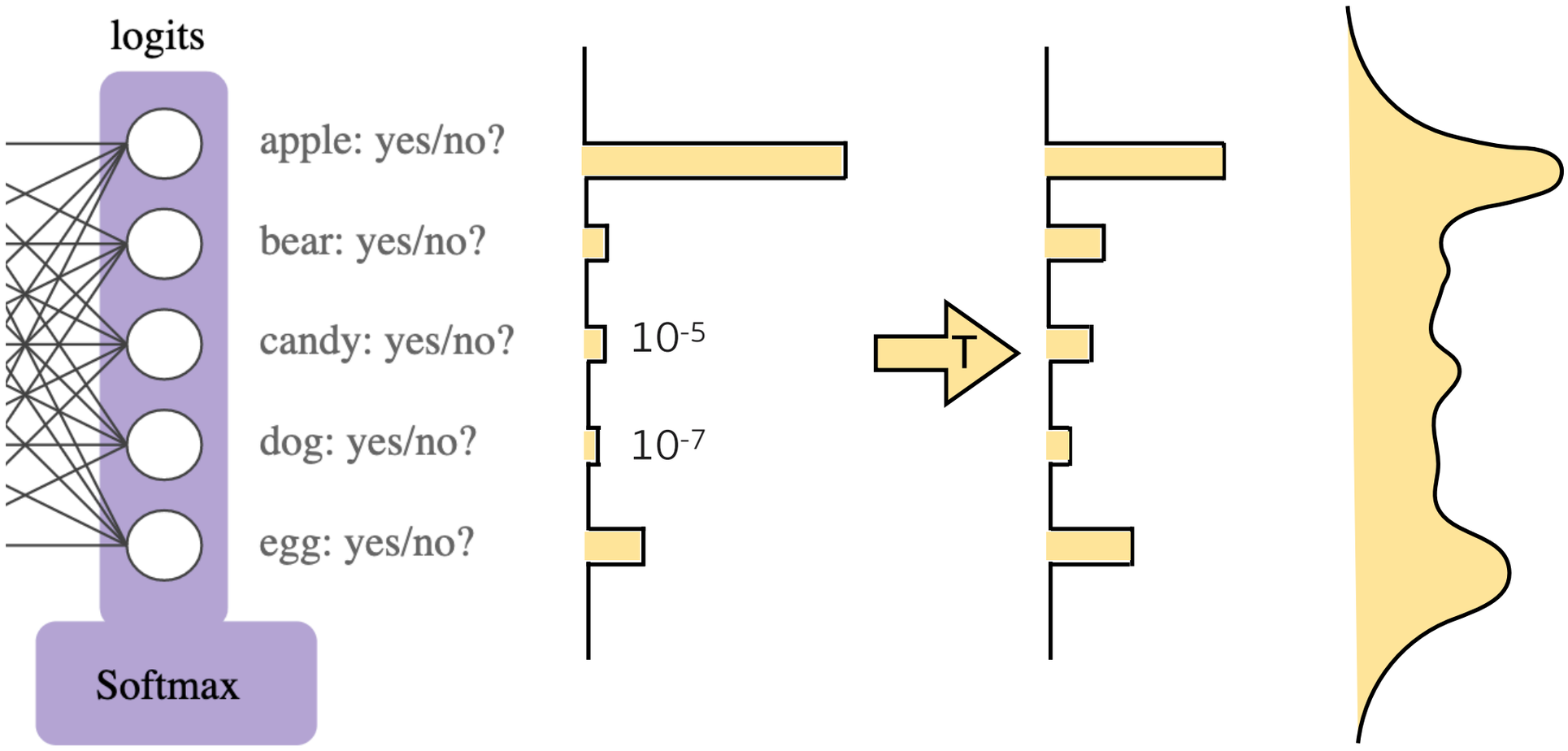
arXiv1503.02531

Keras example

Distillation - raising `Softmax()` temperature **T**, so that the output is suitable soft set of targets

Matching **logits** is a special case of distillation (in the high temperature limit)

`Softmax` = $\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$, where T - temperature parameter



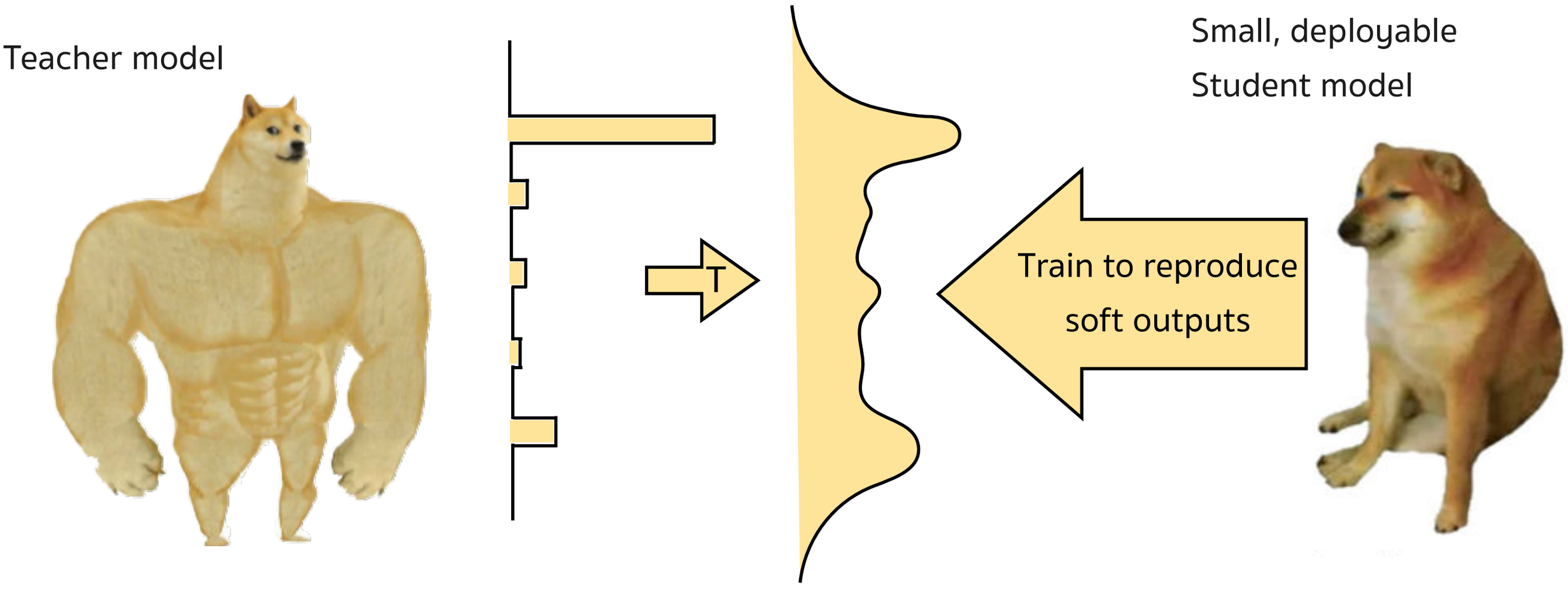
Knowledge Distillation

arXiv1503.02531

KERAS EXAMPLE

Knowledge is transferred to the distilled model by

1. Creating a **transfer set** that is produced by using the pre-trained cumbersome model with a high temperature
2. **Training** the Student on the **transfer set**
3. The **same high temperature** is used when training the distilled model, but after it has been trained it uses a temperature of 1
4. If the **true labels** are known, it can also be used to train the student



Knowledge Distillation

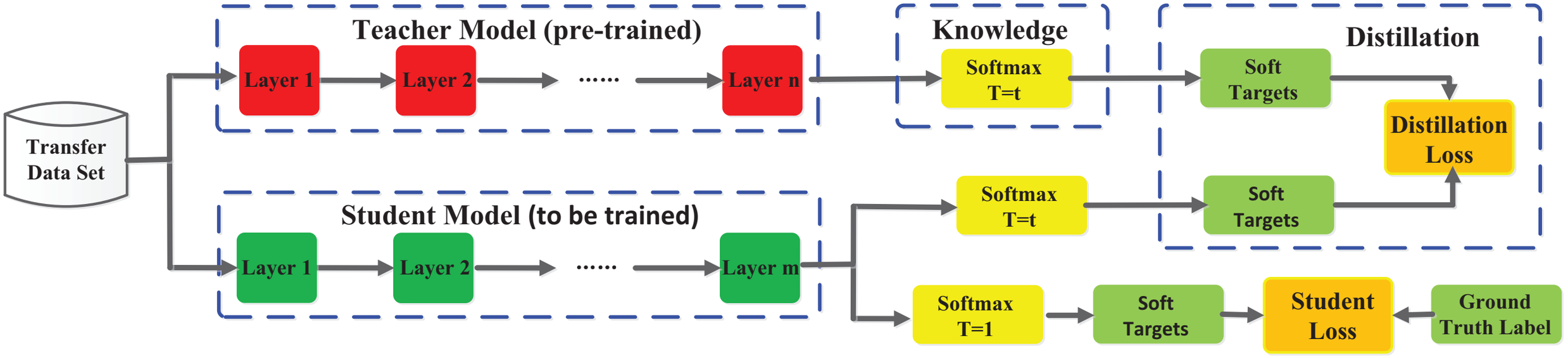
arXiv1503.02531

Keras example

Knowledge is transferred to the distilled model by

1. Creating a **transfer set** that is produced by using the pre-trained cumbersome model with a high temperature
2. **Training** the Student on the **transfer set**
3. The **same high temperature** is used when training the distilled model, but after it has been trained it uses a temperature of 1
4. If the **true labels** are known, it can also be used to train the student

arXiv2006.05525



Results with MNIST

2 hidden layers of

1200 ReLU

Trained on 60k

training cases

67 test errors

arXiv1503.02531

Keras example

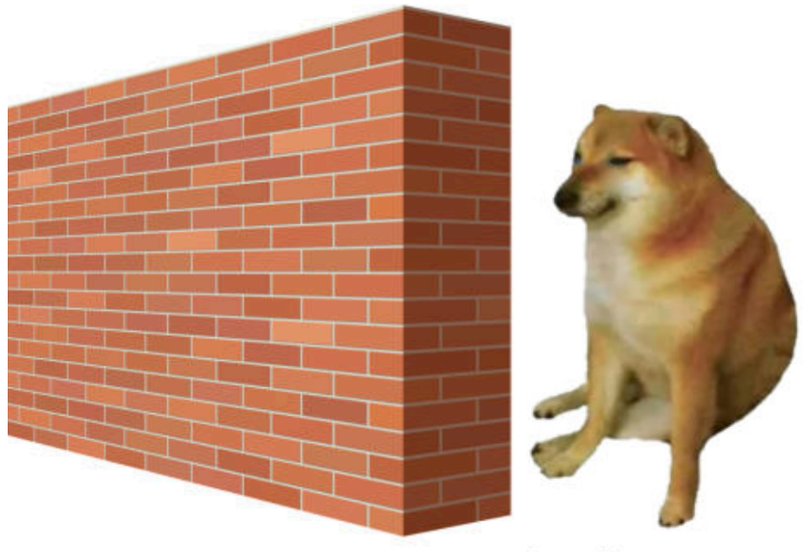
Results with MNIST

arXiv1503.02531

Keras example

2 hidden layers of
1200 ReLU
Trained on 60k
training cases

67 test errors



2 hidden layers of
800 ReLU
Trained on 60k
training cases

146 test errors

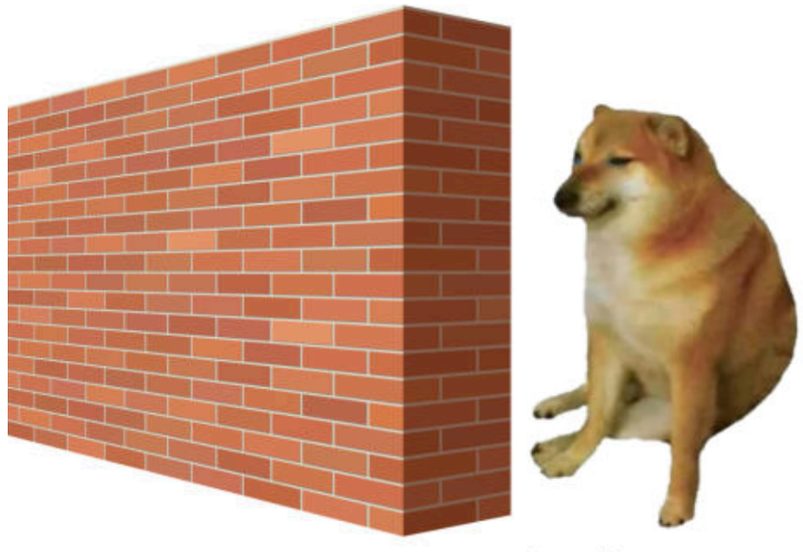
Results with MNIST

arXiv1503.02531

KERAS EXAMPLE

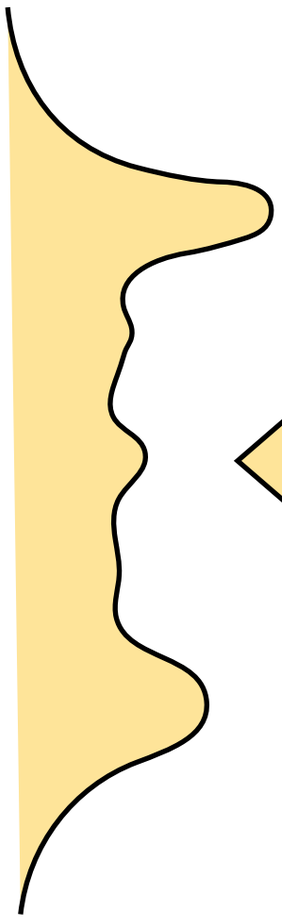
2 hidden layers of
1200 ReLU
Trained on 60k
training cases

67 test errors



2 hidden layers of
800 ReLU
Trained on 60k
training cases

146 test errors



Train to reproduce
soft outputs

2 hidden layers of
800 ReLU
Trained on transfer set
Regularized by adding the task of
matching the soft targets produced by
the large net at `T=20`

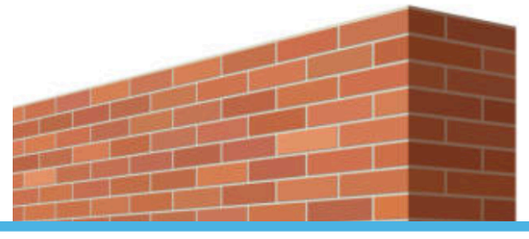
74 test errors

Results with MNIST

arXiv1503.02531

Keras example

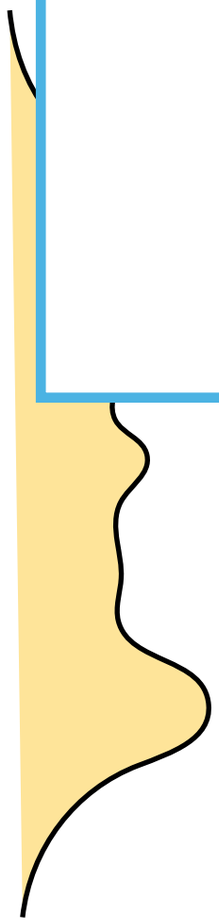
2 hidden layers of
1200 ReLU
Trained on 60k



2 hidden layers of
800 ReLU
Trained on 60k
ses

rors

The authors also experimented with removing some MNIST digits from the transfer set.
The distilled model could still quite well identify them!



Train to reproduce soft outputs

Regularized by adding the task of matching the soft targets produced by the large net at `T=20`

74 test errors

Knowledge Distillation

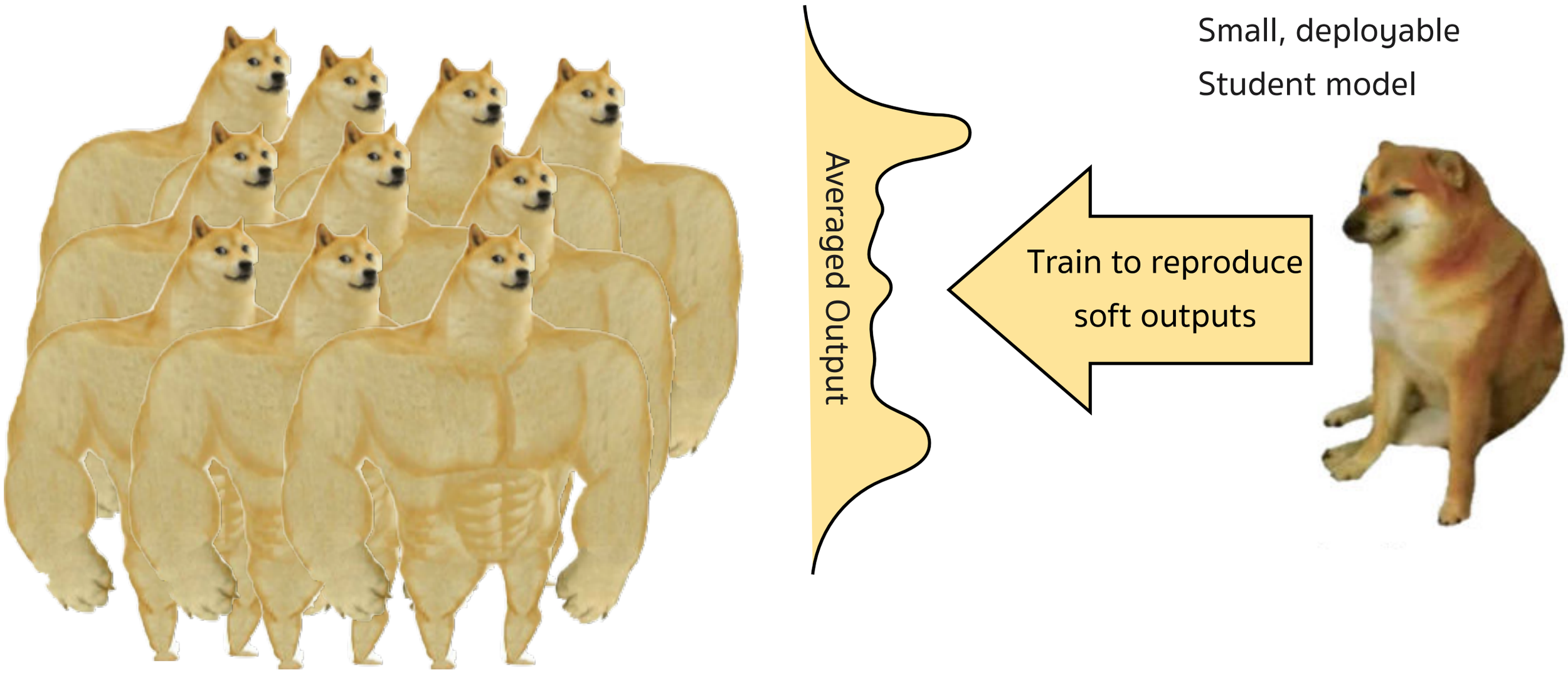
[arXiv1503.02531]

[Keras example]

Knowledge is transferred to the distilled model by

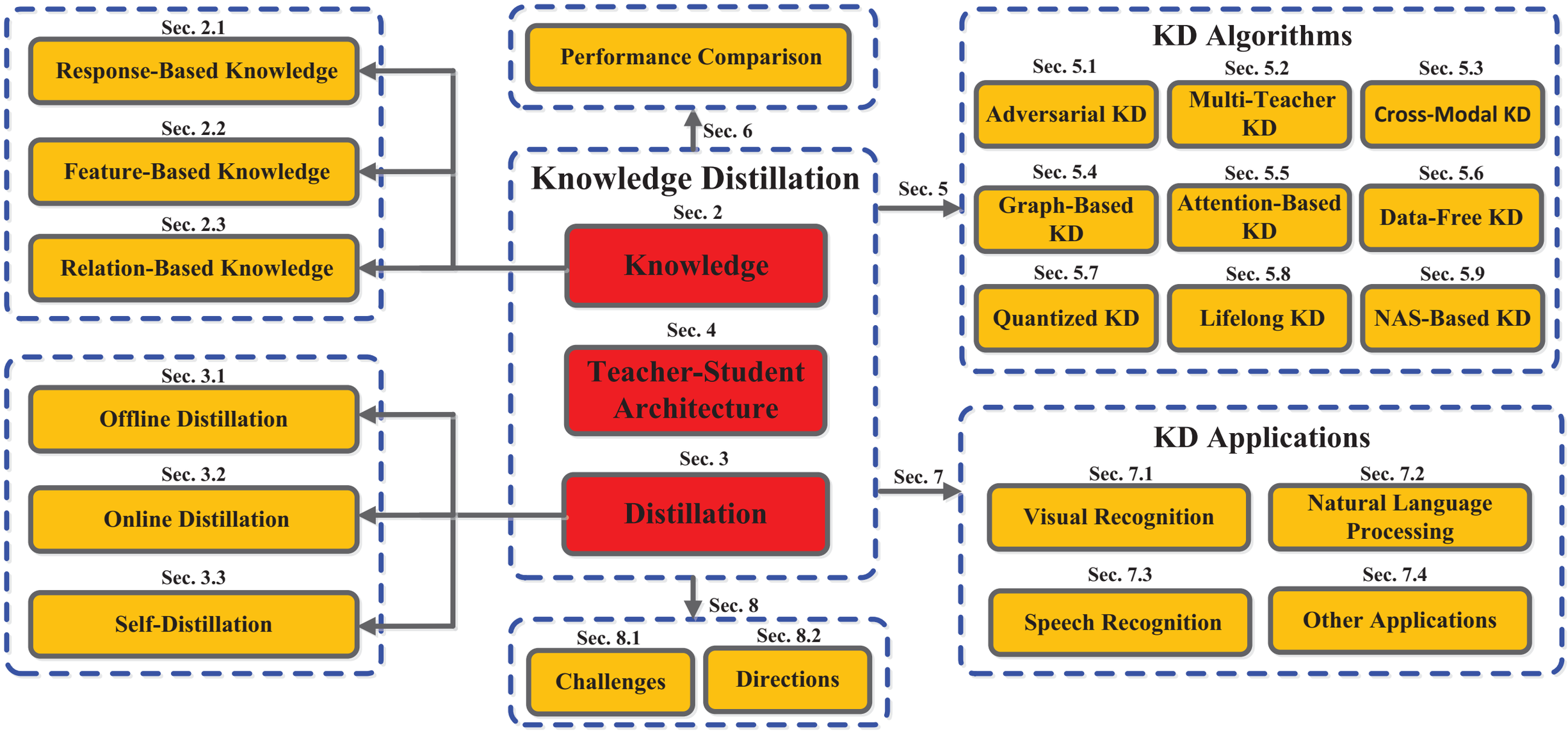
1. Initialising 10 instances of big model
2. Averaging the **soft target** distributions of the ensemble
3. Training the small model to reproduce the **soft target** on a **transfer set**

More than 80% of the improvement in frame classification accuracy achieved by using an ensemble of 10 models is transferred to the distilled model



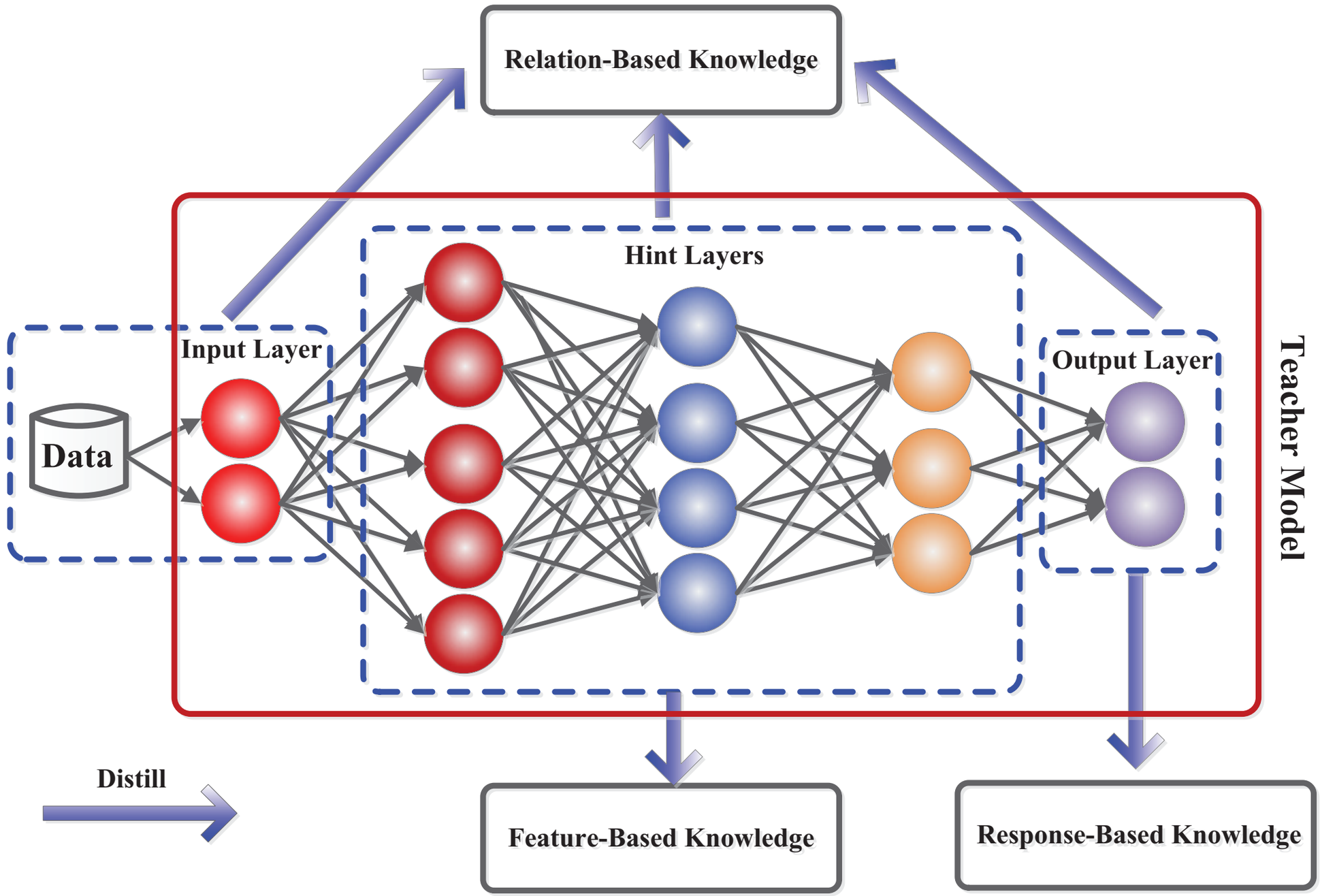
arXiv2006.05525

The schematic structure of knowledge distillation and the relationship between the adjacent sections



ARXIV2006.05525

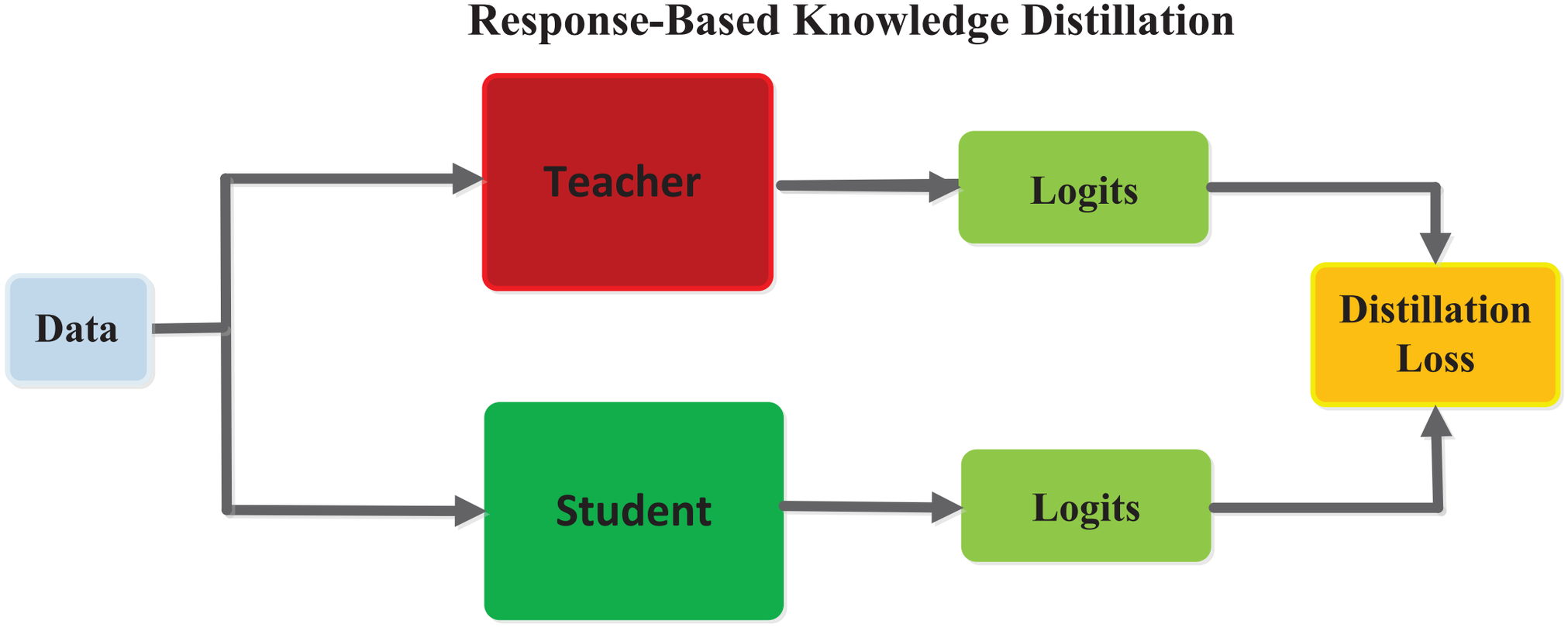
The schematic structure of knowledge types



Response-Based Knowledge

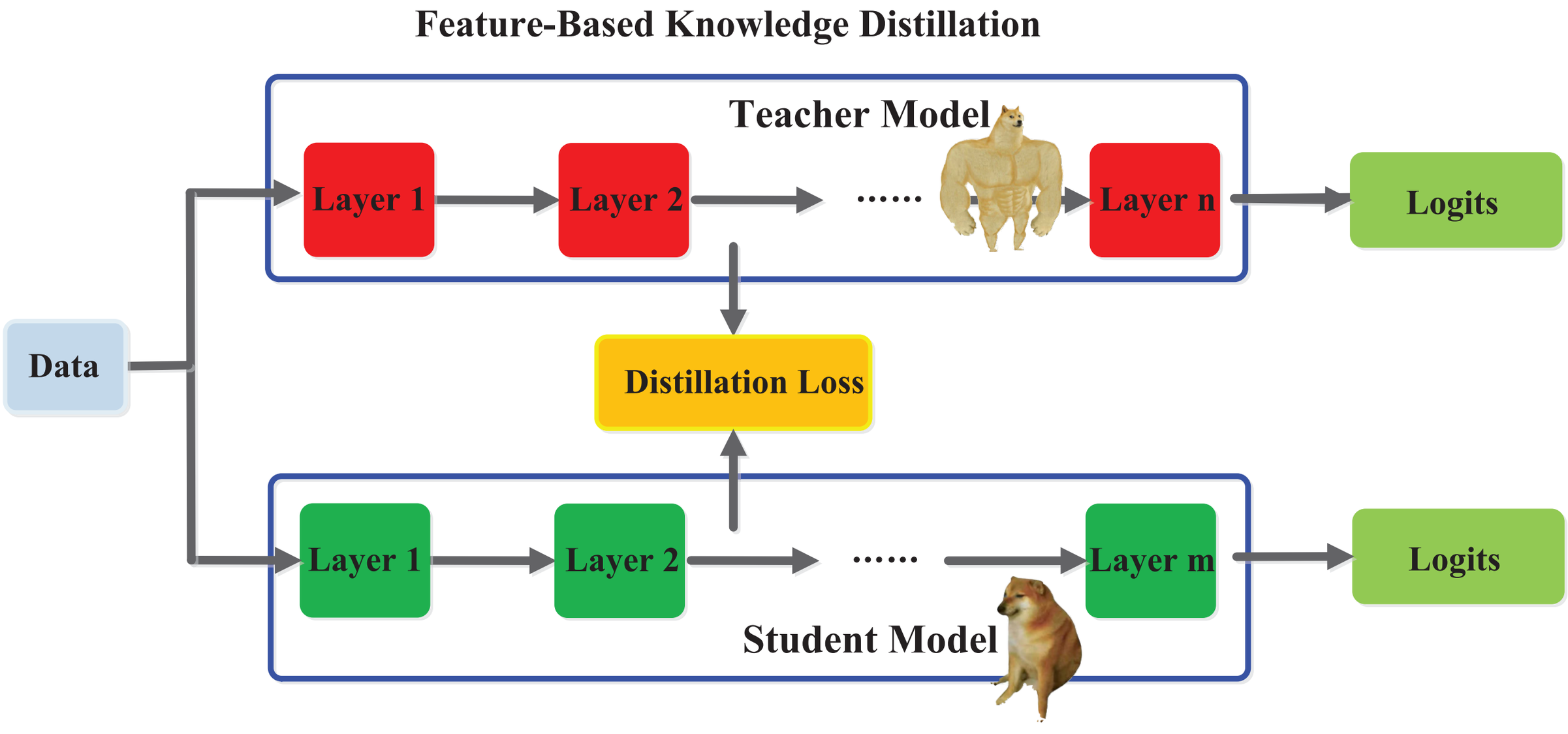
arXiv2006.05525

- Refers to the neural response of the last output layer of the teacher model
- The main idea is to directly mimic the final prediction of the teacher model
- The response-based knowledge distillation is simple yet effective for model compression, and has been widely used in different tasks and applications



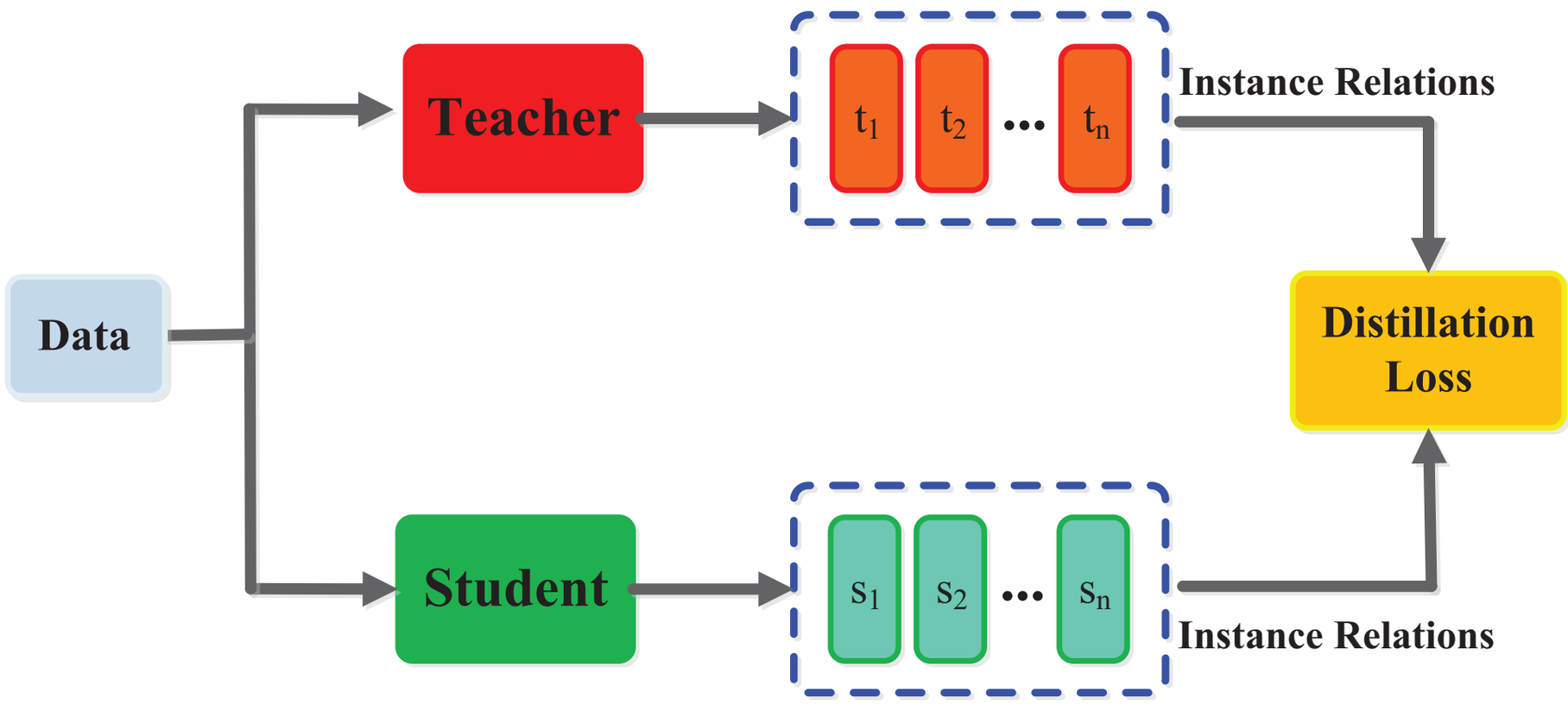
ARXIV2006.05525

- Both the output of the last layer and the output of intermediate layers, i.e., feature maps, can be used as the knowledge to supervise the training of the student model
- An “attention map” from the original feature maps can be used to express knowledge



ARXIV2006.05525

- The individual soft targets of a teacher are directly distilled into student
- The distilled knowledge contains not only feature information but also mutual relations of data samples
- The transferred knowledge in instance relationship graph contains instance features, instance relationships and the feature space transformation cross layer



arXiv2006.05525

Offline Distillation

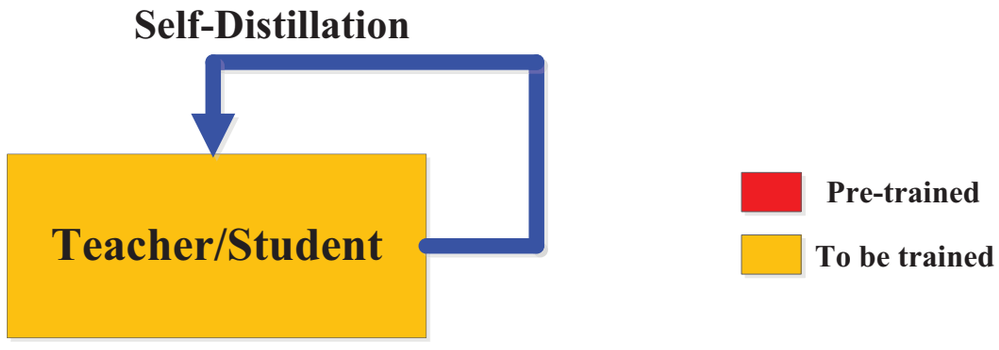
the knowledge is transferred from a pre-trained teacher model into a student model

Online Distillation

is a one-phase end-to-end training scheme with efficient parallel computing, usually all models are of the same size

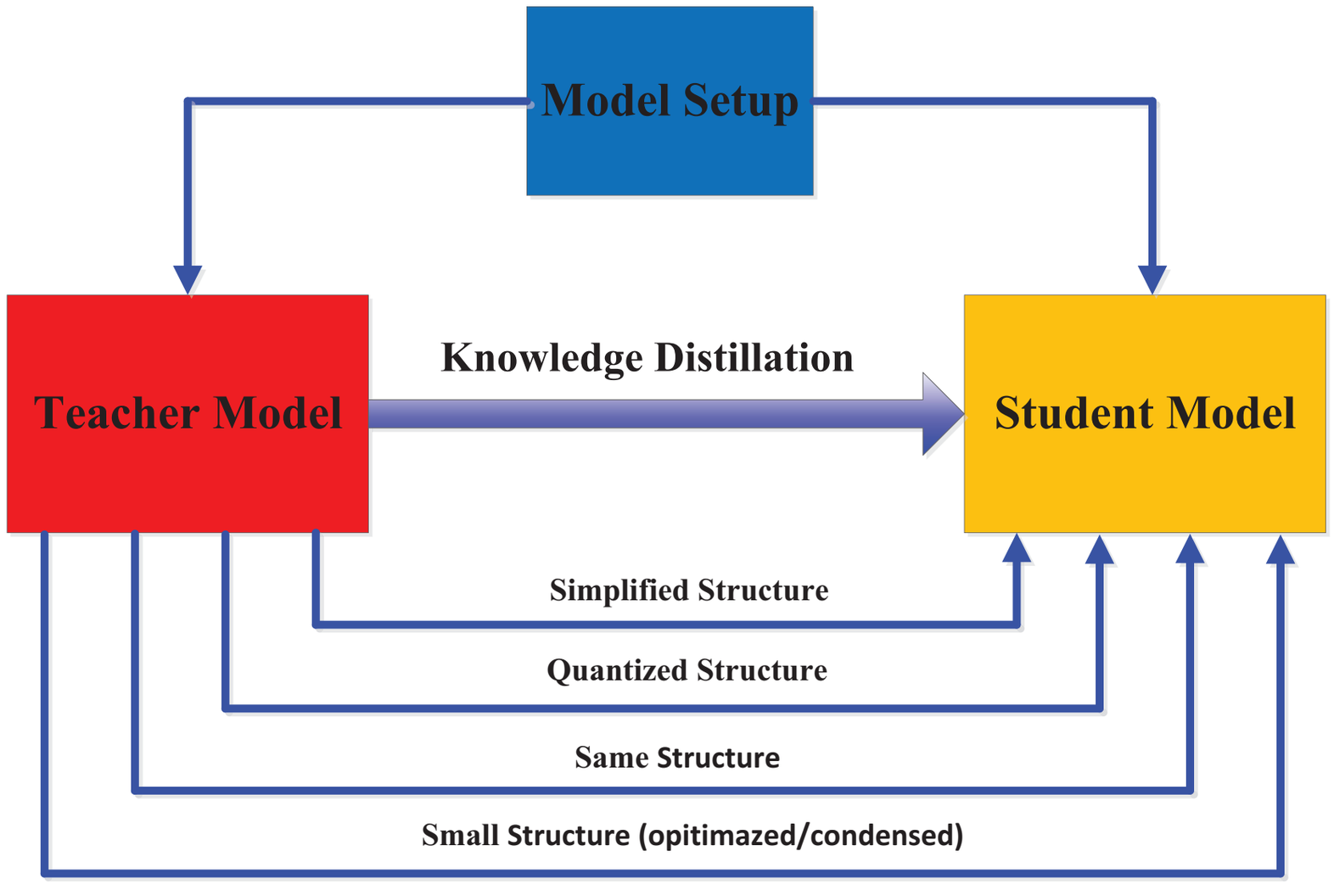
Self-Distillation

knowledge from the deeper sections of the network is distilled into its shallow sections



ARXIV2006.05525

The idea of a neural architecture search in knowledge distillation, i.e., a joint search of student structure and knowledge transfer under the guidance of the teacher model, will be an interesting subject of future study



arXiv2006.05525

If you are interested in using KD, consider the following:

- Teacher-Student architectures
- Distillation schemes (**offline**/online/self)
- Knowledge type (**response**/feature/relation-based)
- Knowledge distillation algorithm to use

Potential Application of KD For hls4ml

- In hls4ml large application of NN in L1
- How to use KD for autoencoders?
- KD for unsupervised AD arXiv1911.02357
- @Hassan is working on GarNet autoencoder for HL-LHC L1 anomaly detection. First synthesis results give O(10μs) latency, while we need O(1μs)
- Can be useful for jet classification?
- Can we use KD to teach the knowledge of trained GarNet to a smaller (MLP?) network? Do we want to do that? (Need to show that MLP alone is worse than GarNet)