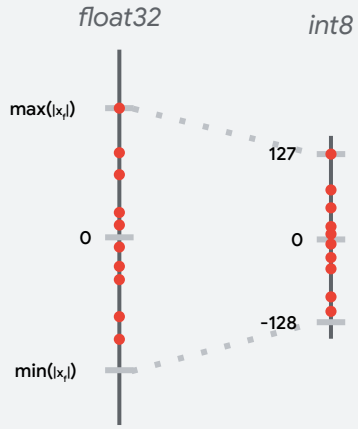
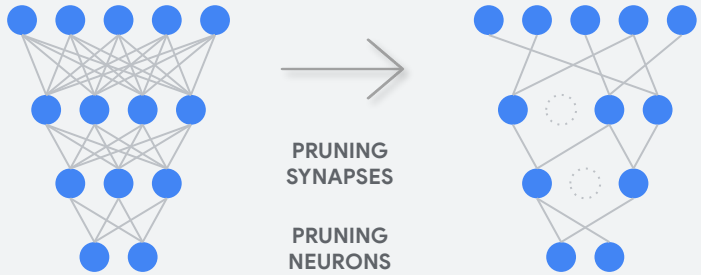
Knowledge Distillation

Optimizations So Far

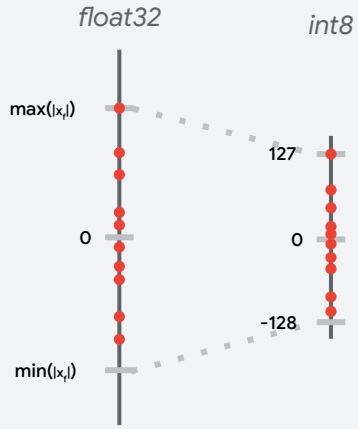
- Pruning, **removing weights or activations** close to zero
- Model Quantization, **low-precision arithmetic**

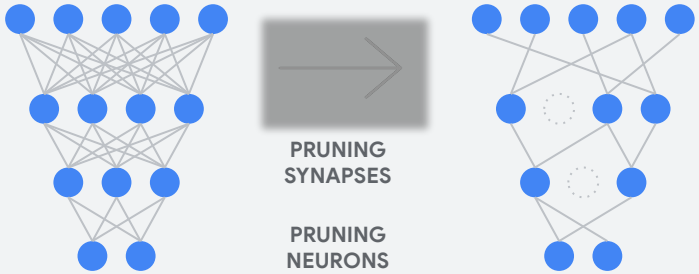




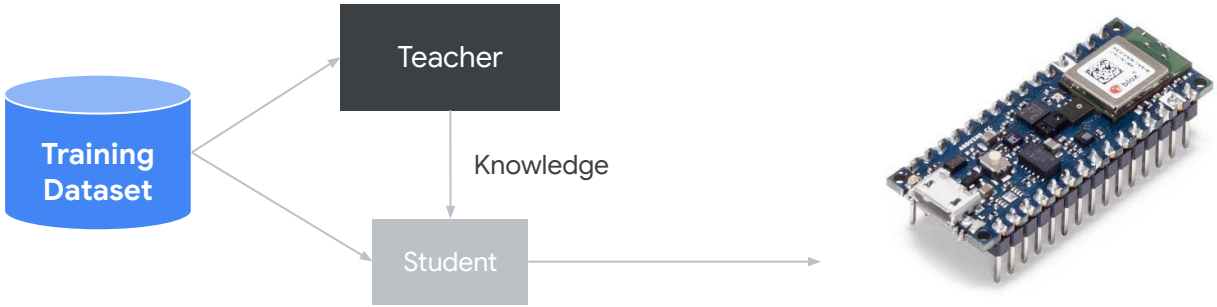
Optimizations So Far

- Pruning, **removing weights or activations** close to zero
- Model Quantization, **low-precision arithmetic**
- Knowledge Distillation **distills its knowledge**



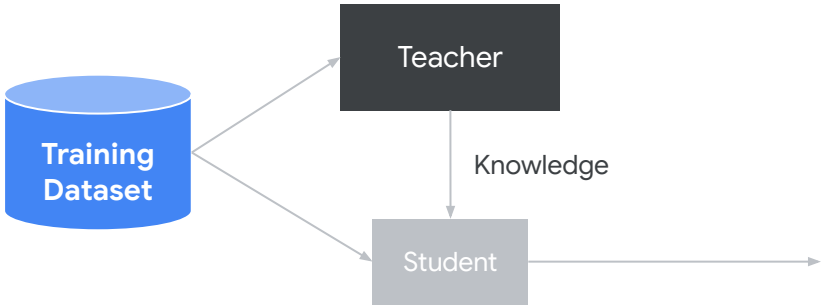


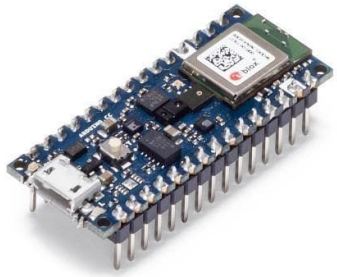
Knowledge Distillation



Knowledge Distillation

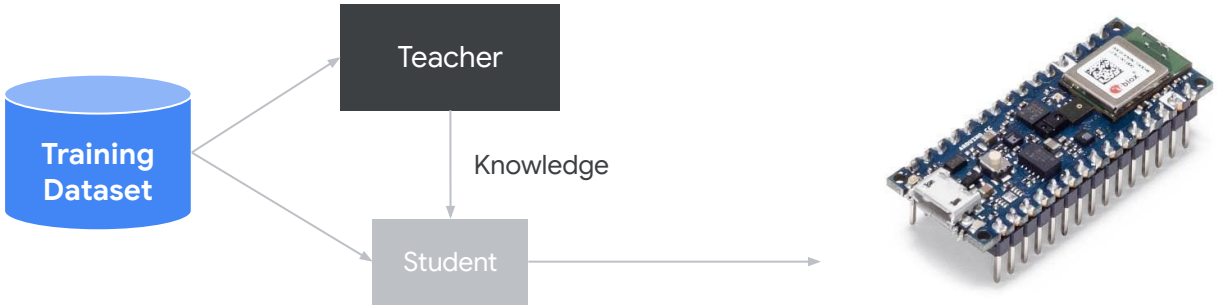
- Knowledge Distillation where a large complex model **(teacher)** distills its knowledge and passes it to train a smaller network **(student)**





Knowledge Distillation

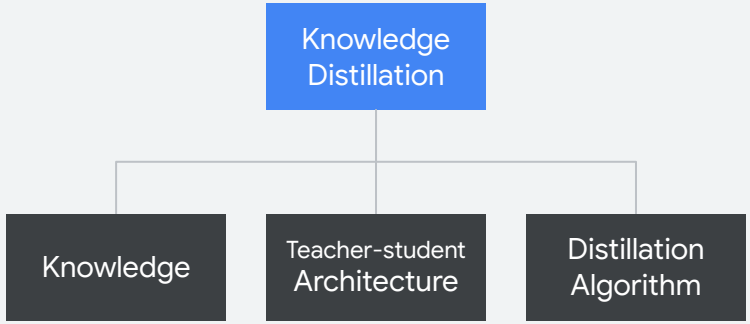
- Knowledge Distillation where a large complex model **(teacher)** distills its knowledge and passes it to train a smaller network **(student)**



- The student network is trained to match the **larger network's prediction** and **the distribution of the teacher's network**.

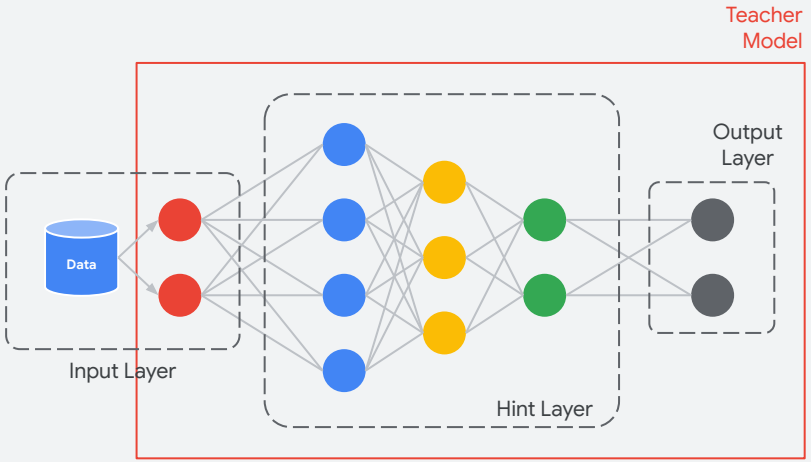
KD: Building Blocks

A knowledge distillation system consists of **three principal components**



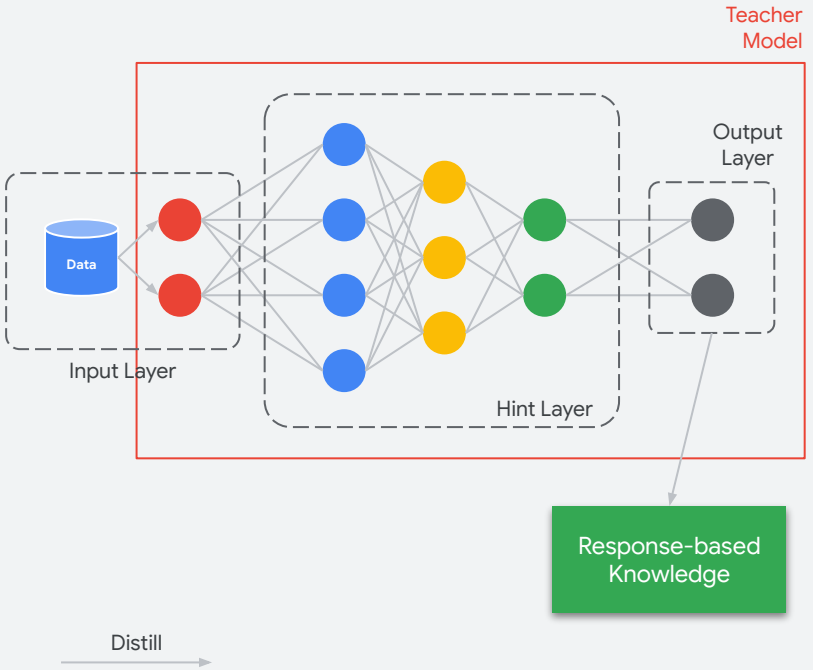
Knowledge Types

1. **Response**-based knowledge
2. **Feature**-based knowledge
3. **Relation**-based knowledge



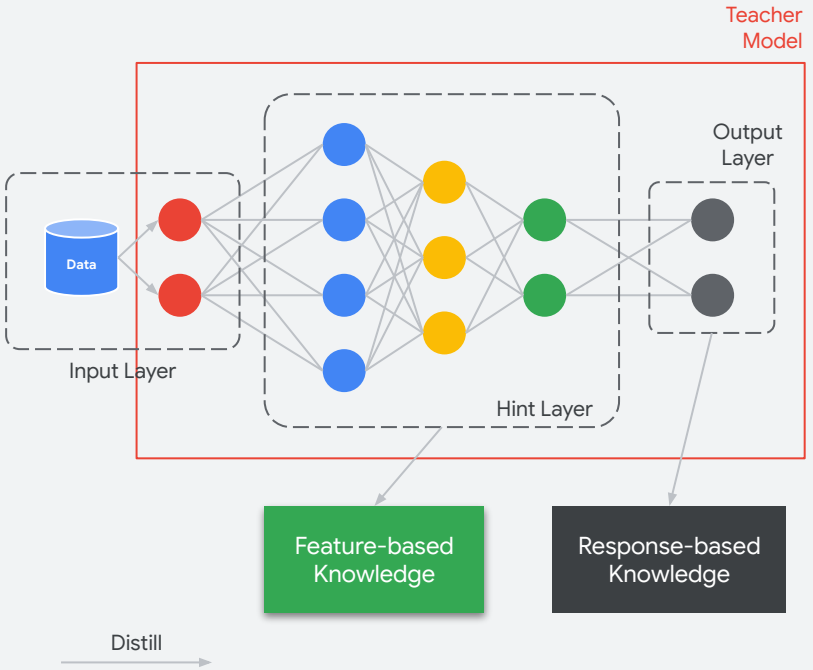
Knowledge Types

1. **Response**-based knowledge
2. **Feature**-based knowledge
3. **Relation**-based knowledge



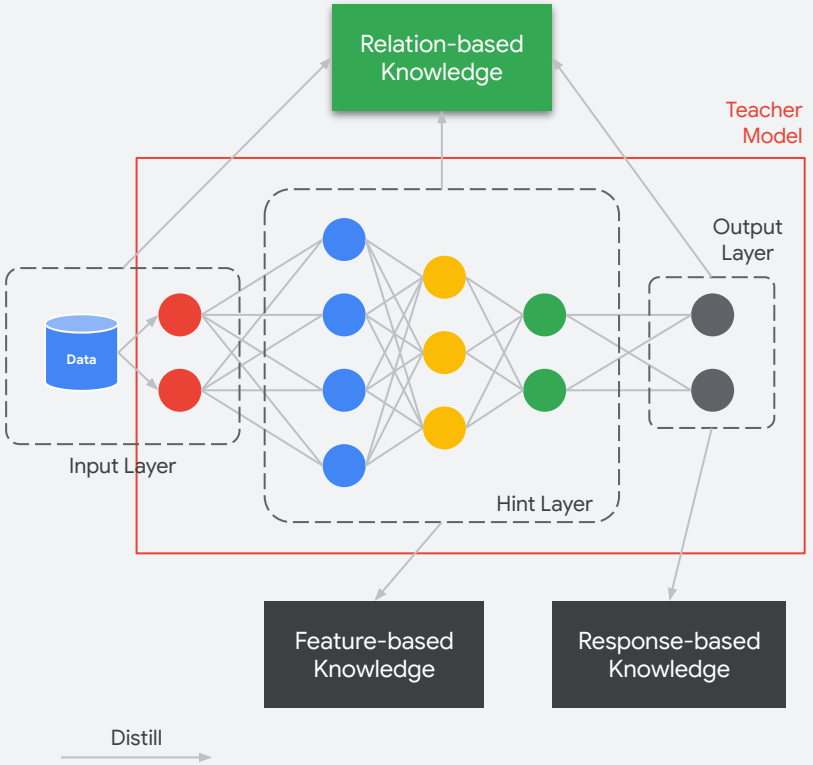
Knowledge Types

1. **Response**-based knowledge
2. **Feature**-based knowledge
3. **Relation**-based knowledge



Knowledge Types

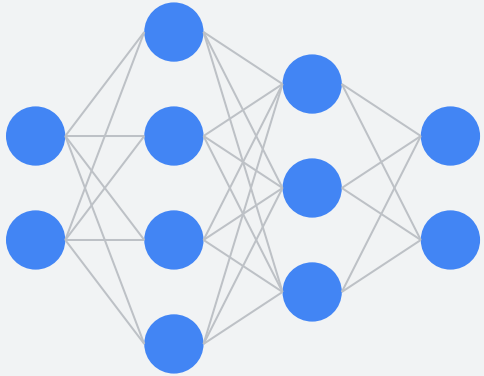
1. **Response**-based knowledge
2. **Feature**-based knowledge
3. **Relation**-based knowledge



Student-Teacher Architecture

Need to determine what the neural network architecture looks like for the **student network**

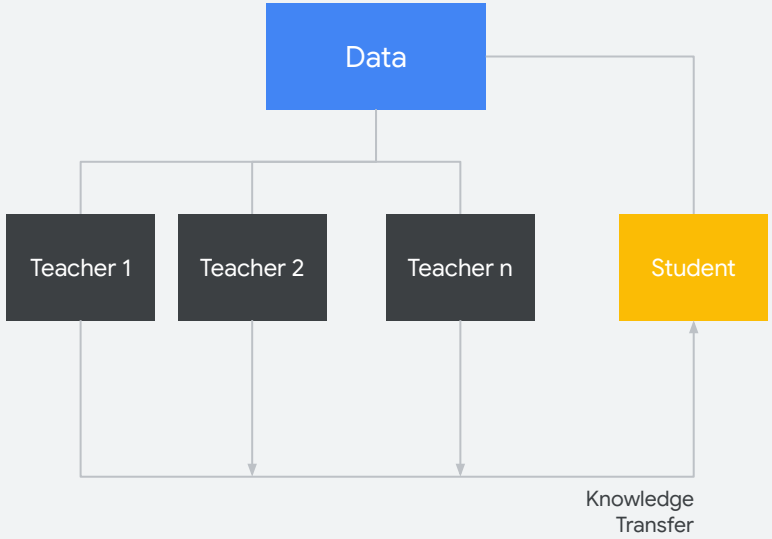
- Size
- Latency
- Efficiency
- ...



Knowledge Distillation Algorithm

Many different ways to **teach** the student

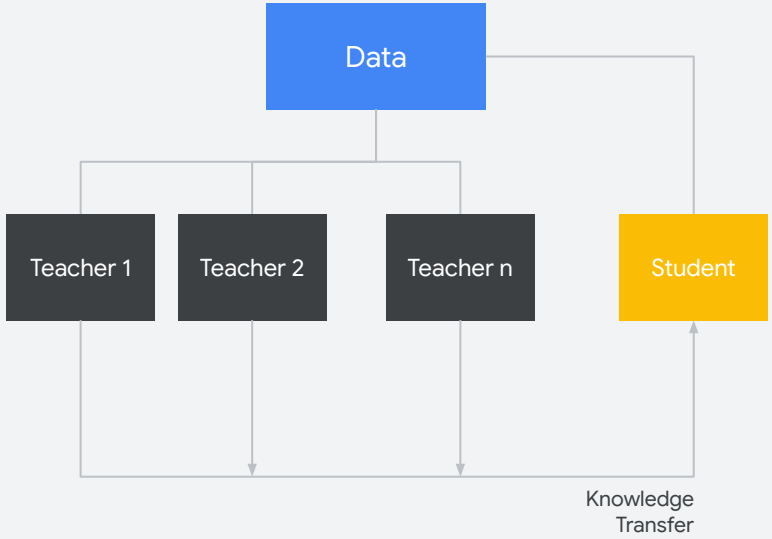
- Ensemble distillation



Knowledge Distillation Algorithm

Many different ways to **teach** the student

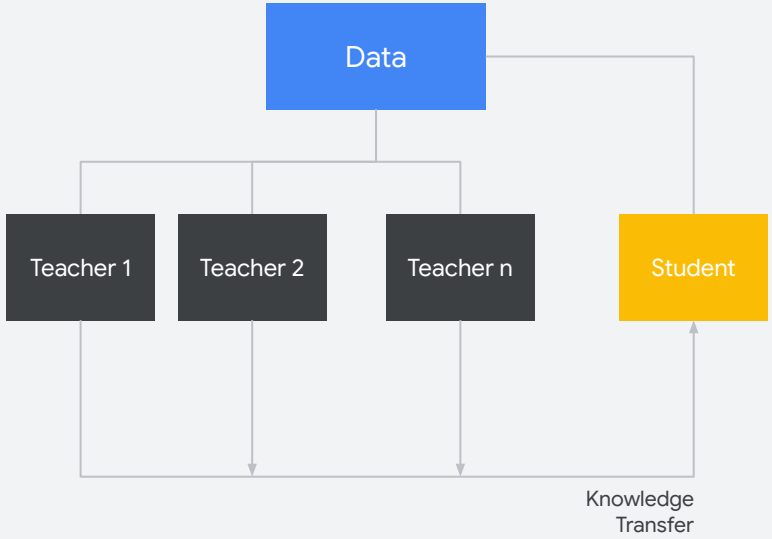
- Ensemble distillation
- Quantized distillation



Knowledge Distillation Algorithm

Many different ways to **teach** the student

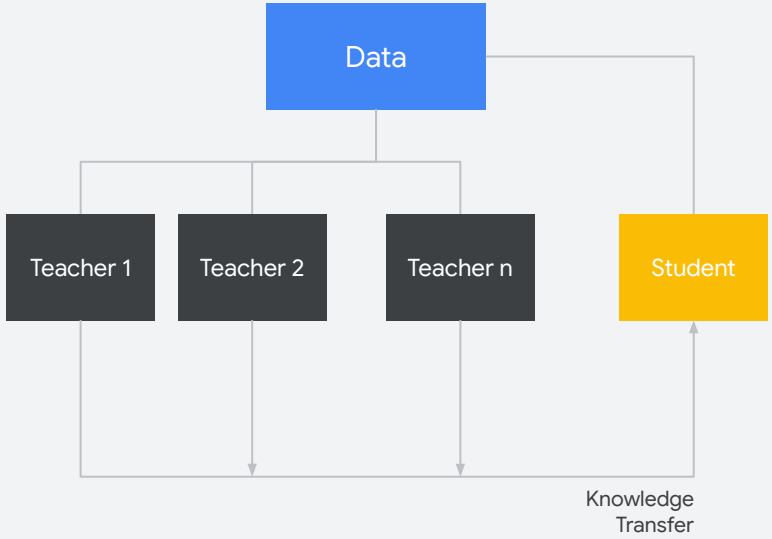
- Ensemble distillation
- Quantized distillation
- NAS distillation

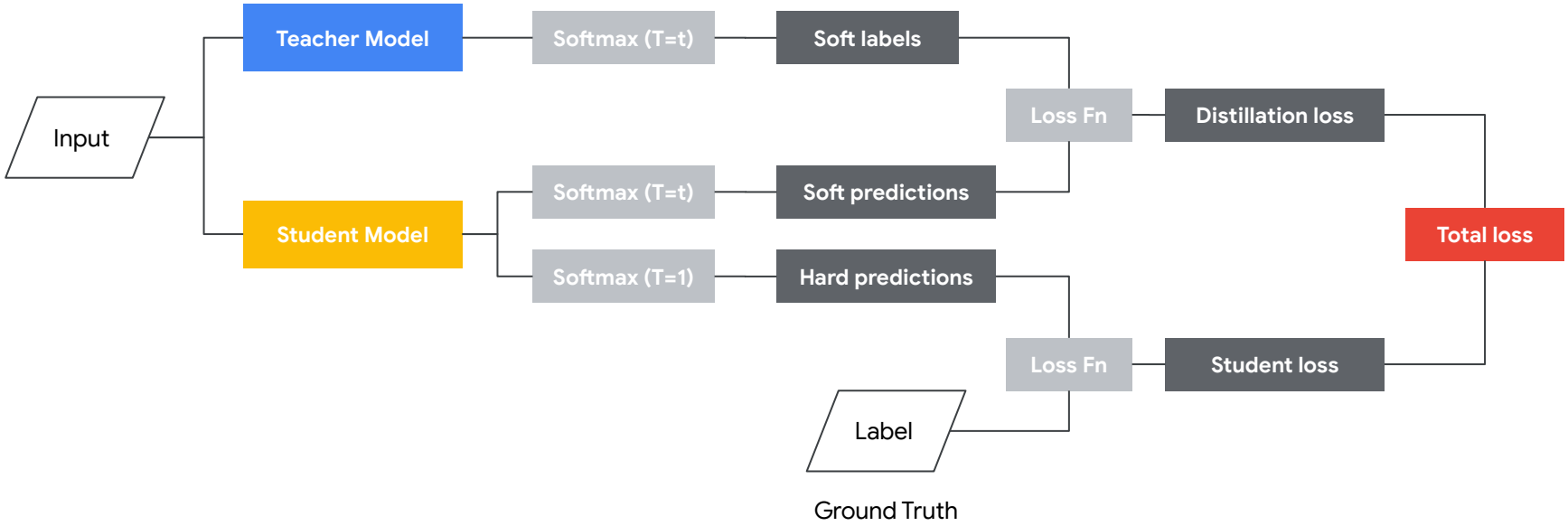


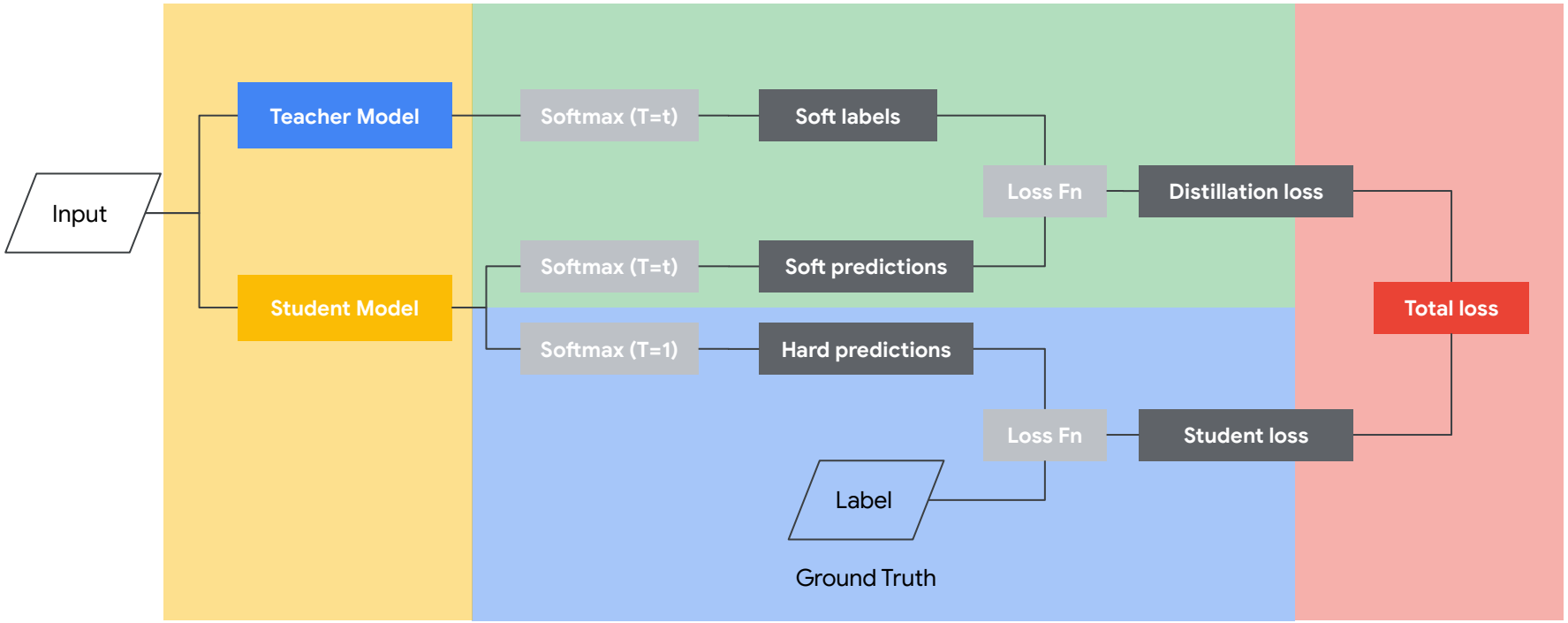
Knowledge Distillation Algorithm

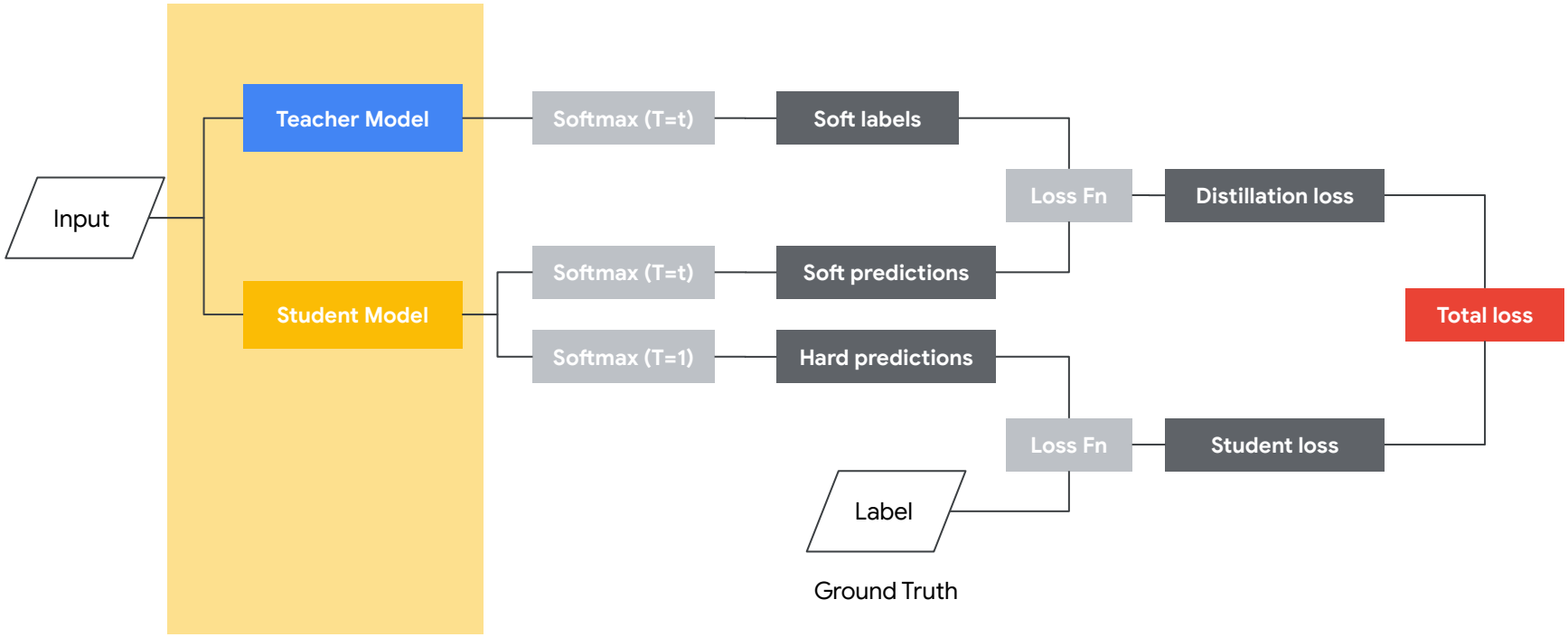
Many different ways to **teach** the student

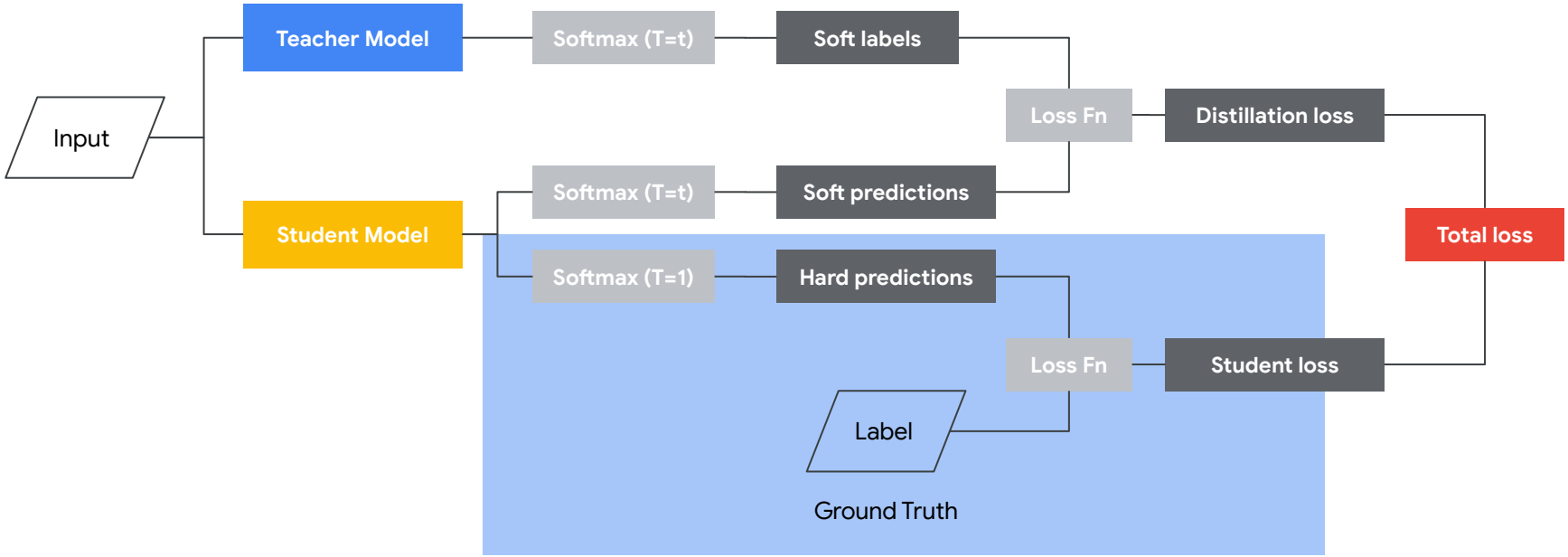
- Ensemble distillation
- Quantized distillation
- NAS distillation
- Attention distillation

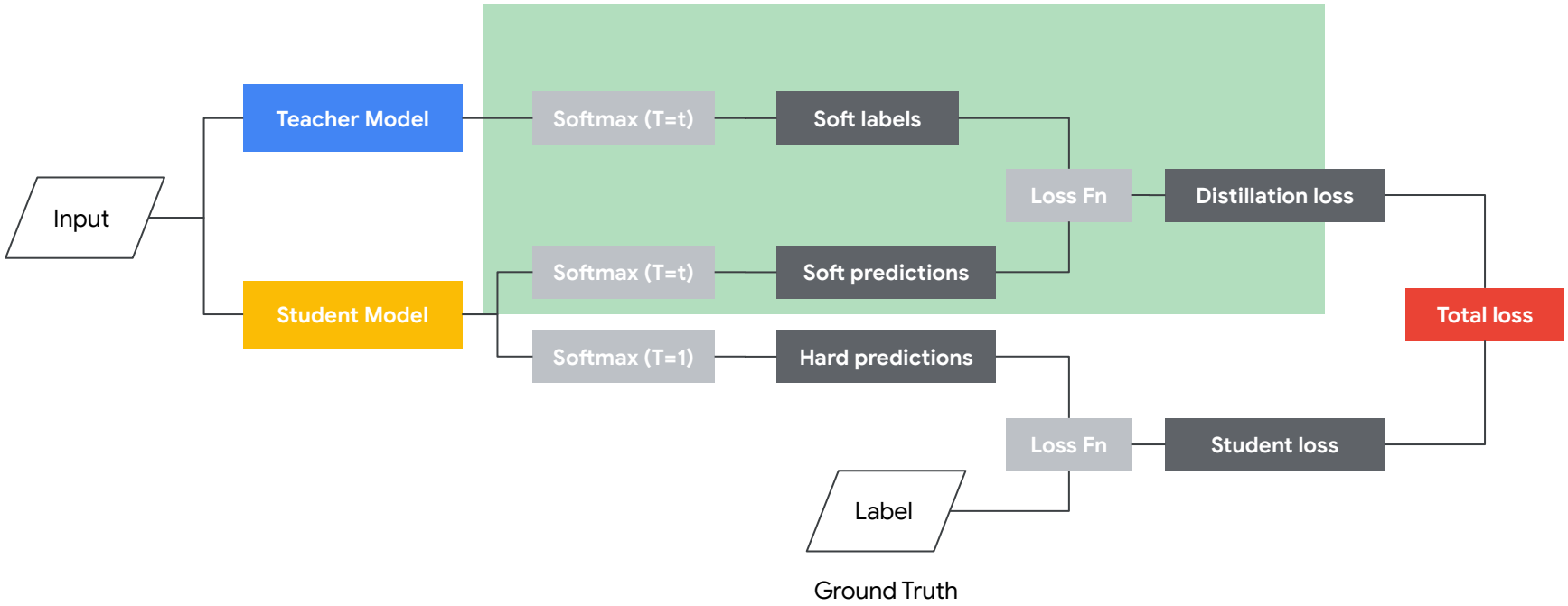


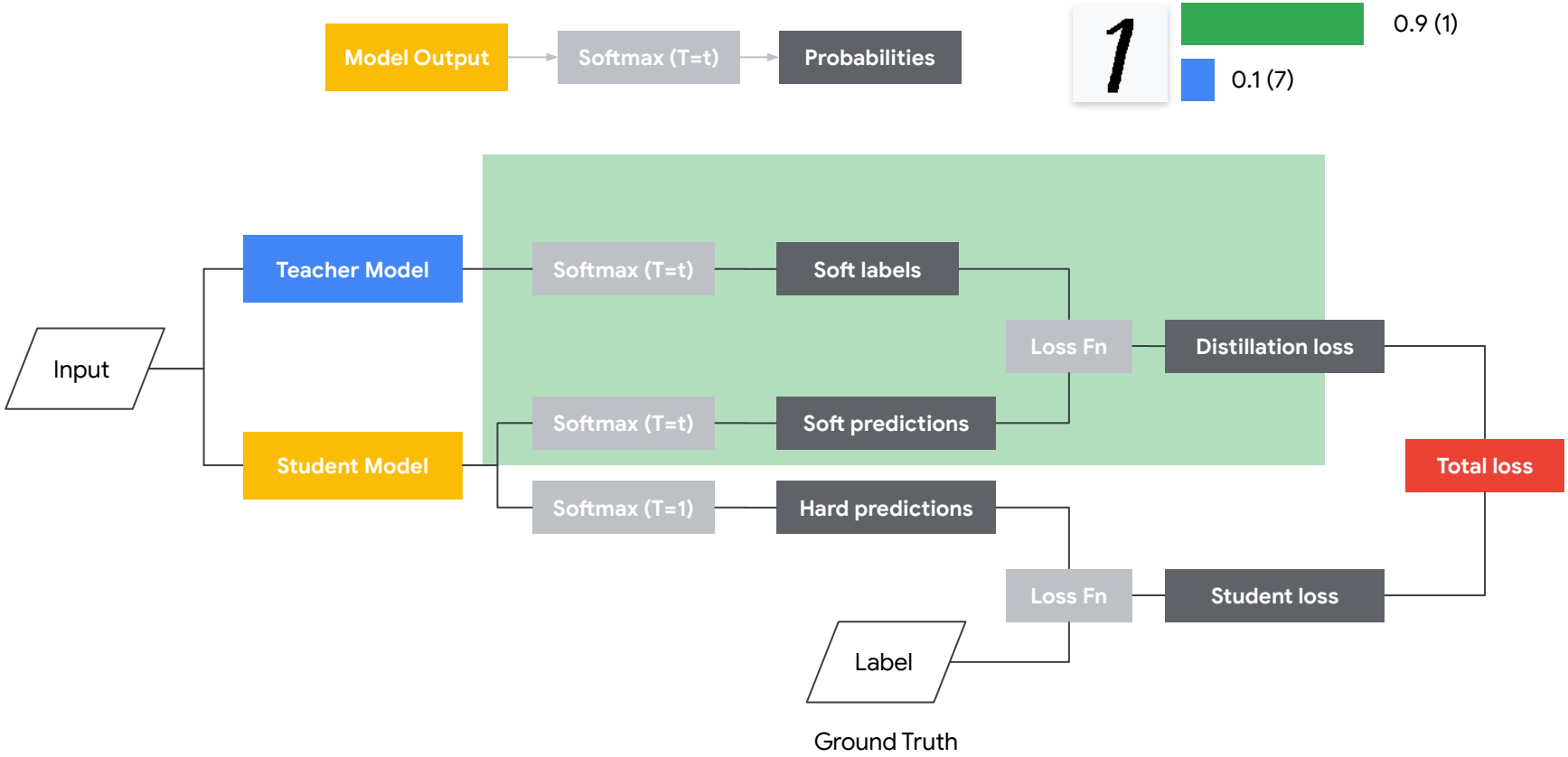


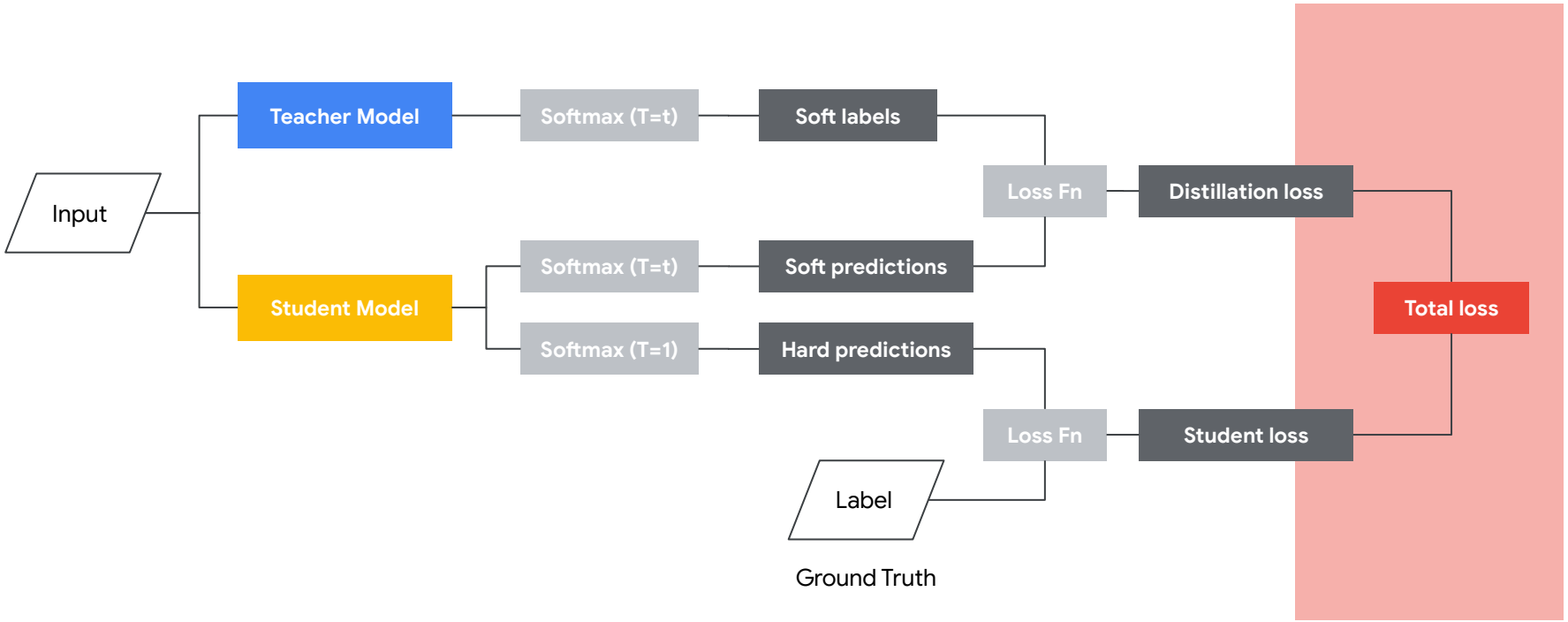










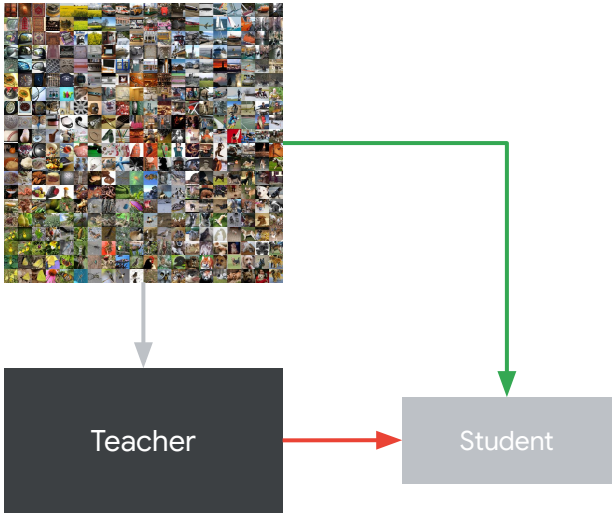


Knowledge Distillation Limitations

A large difference in model **Capacity** and/or **Architecture** between the teacher and student can limit the effectiveness of Knowledge Distillation

Knowledge Distillation Limitations

Knowledge distillation can perform **worse** than training from **scratch** on large challenging datasets (E.g. Imagenet)



Knowledge Distillation Limitations

Knowledge distillation can perform **worse** than training from **scratch** on large challenging datasets (E.g. Imagenet)