HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

# Clean-IT: Towards Sustainable Digital Technologies
# Model Compression using Knowledge Distillation

**Christian Bartz**

PhD Student

# What is Knowledge Distillation?

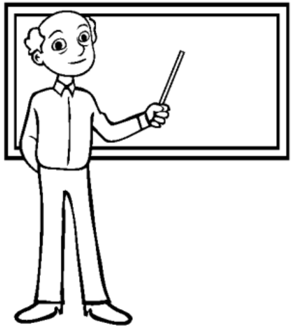
- knowledge distillation is a method for model compression
- first introduced by Hinton et al. in 2015 [1]

[1] - Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

# What is Knowledge Distillation?

- knowledge distillation is a method for model compression
- first introduced by Hinton et al. in 2015 [1]
- knowledge of a teacher model is transferred/distilled into a student model

Teacher



pixy.org, CC BY-NC-ND 4.0

Student

flyclipart.com, CC BY-NC-ND 4.0

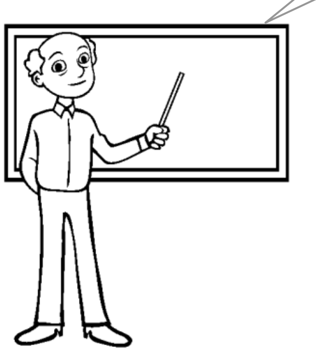
[1] - Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).

# What is Knowledge Distillation?

- knowledge distillation is a method for model compression
- first introduced by Hinton et al. in 2015 [1]
- knowledge of a teacher model is transferred/distilled into a student model

This is a dog

Teacher



pixy.org, CC BY-NC-ND 4.0

Biser Yanev, CC BY-SA 4.0 via Wikimedia Commons

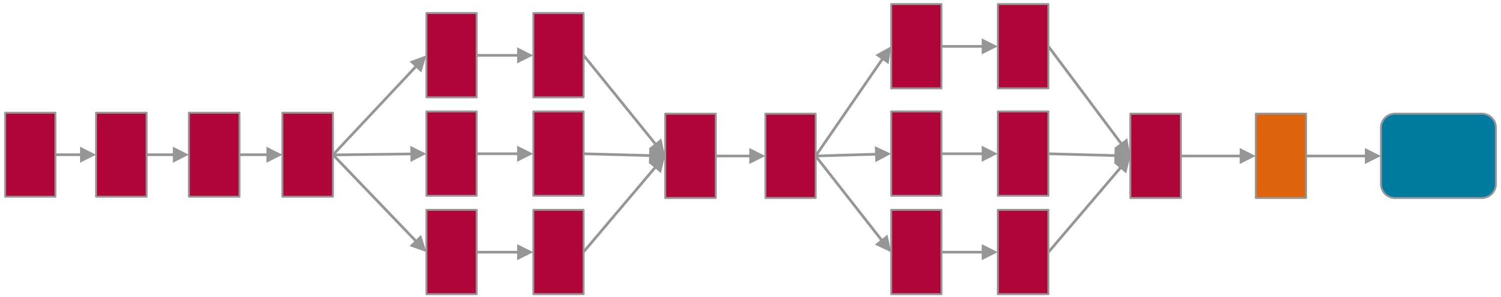
Okay! Got it!

Student

flyclipart.com, CC BY-NC-ND 4.0

# Knowledge Distillation in Neural Networks

Teacher Network



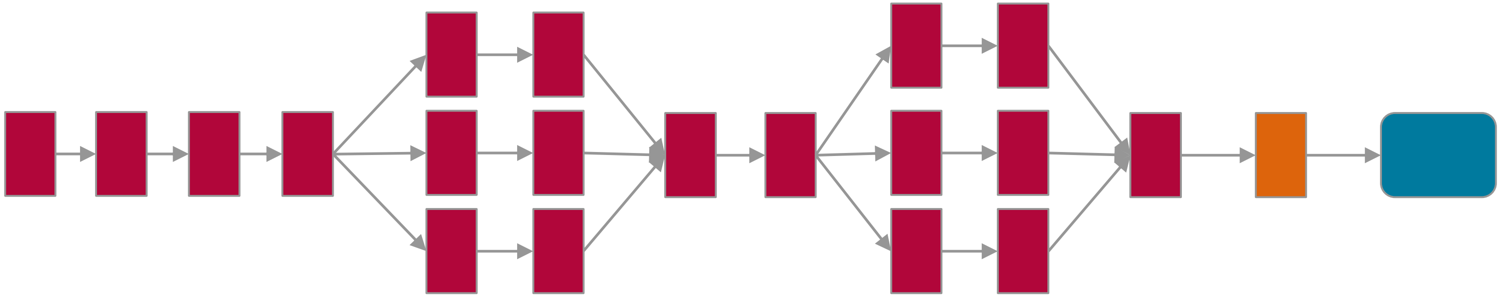
🟥 Convolution

🟧 Fully Connected

🟦 Softmax

# Knowledge Distillation in Neural Networks

## Teacher Network



🟥 Convolution

🟧 Fully Connected

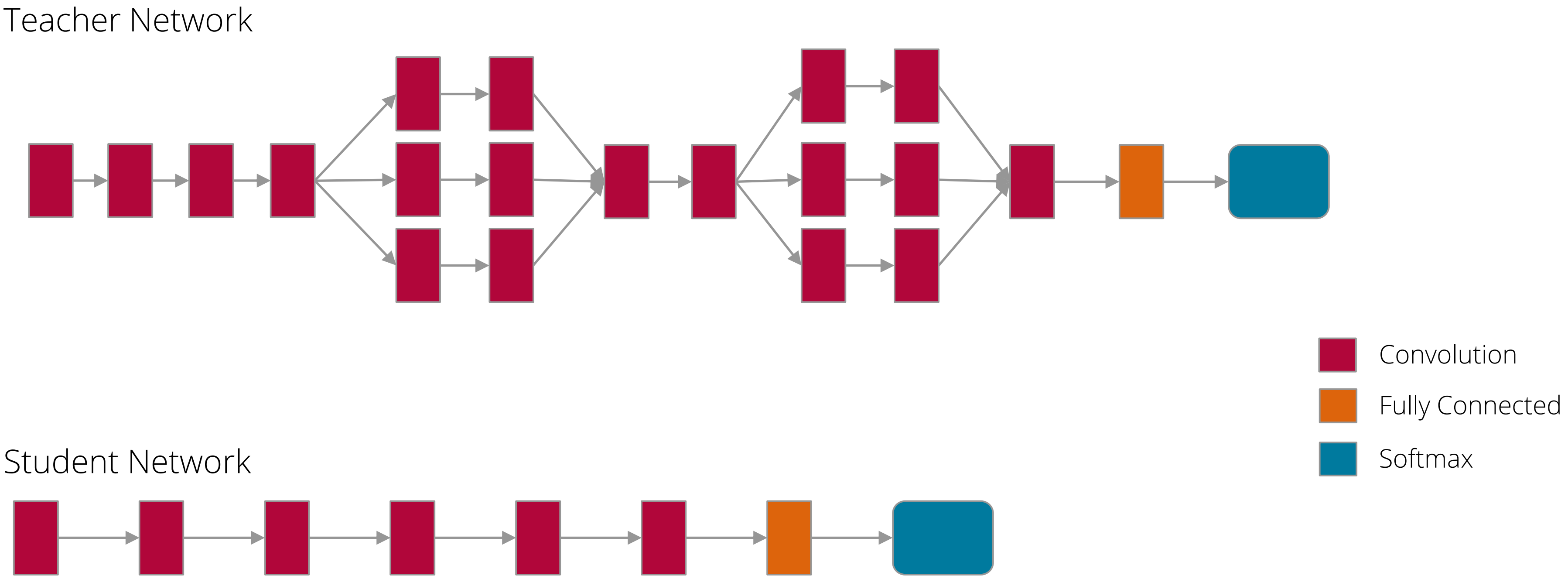
🟦 Softmax

- trained by `hard` labels and softmax cross entropy

Biser Yanev, CC BY-SA 4.0 via Wikimedia Commons

| | | |
|---|---|---|
| dog | 0 | [1, 0, 0, 0] |
| cat | 1 | [0, 1, 0, 0] |
| car | 2 | [0, 0, 1, 0] |
| ship | 3 | [0, 0, 0, 1] |

# Knowledge Distillation in Neural Networks



Teacher Network

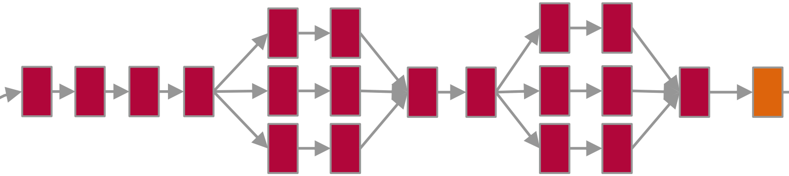
Student Network

🟥 Convolution

🟧 Fully Connected

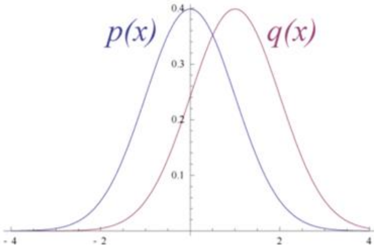
🟦 Softmax

# Training the Student

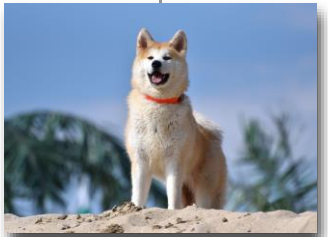


[5, -3, -5, -7]

softmax cross entropy loss



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

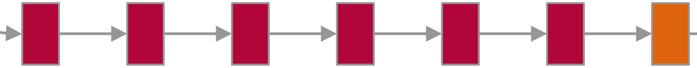


Biser Yanev, CC BY-SA 4.0 via Wikimedia Commons

🟥 Convolution

🟧 Fully Connected

🟦 Softmax



[-1, 3, 2, -3]

# Case Study
ImageNet Training

| Method | Parameters | Model Size | Top 1 Accuracy |
|---|---|---|---|
| Teacher (ResNet-152) | 60,344,232 | 244 MB | 77.98% |

Tang, Jiaxi, et al. "Understanding and Improving Knowledge Distillation." *arXiv preprint arXiv:2002.03532* (2020).

# Case Study
# ImageNet Training

| Method | Parameters | Model Size | Top 1 Accuracy |
|---|---|---|---|
| Teacher (ResNet-152) | 60,344,232 | 244 MB | 77.98% |
| Student (ResNet-50) | 25,610,216 | 104 MB | 76.32% |

Tang, Jiaxi, et al. "Understanding and Improving Knowledge Distillation." *arXiv preprint arXiv:2002.03532* (2020).

# Case Study
# ImageNet Training

| Method | Parameters | Model Size | Top 1 Accuracy |
|---|---|---|---|
| Teacher (ResNet-152) | 60,344,232 | 244 MB | 77.98% |
| Student (ResNet-50) | 25,610,216 | 104 MB | 76.32% |
| Knowledge Distillation | 25,610,216 | 104 MB | 77.75% |

Tang, Jiaxi, et al. "Understanding and Improving Knowledge Distillation." *arXiv preprint arXiv:2002.03532* (2020).

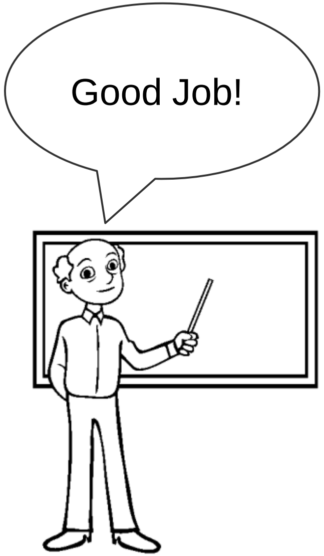
# Why does it work?

- models are often overparameterized
- soft targets contain more information about data than hard labels
  - i.e. a dog is more related to a cat, than a dog to a car
- soft targets also have less gradient variance

→ smoother and easier training

# Conclusion

- model distillation is a method for model compression
  - utilizes fact that large models are often overparameterized
- model distillation involves teacher and student
  - student learns based on soft labels produced by the teacher
- model distillation reduces energy usage of AI
- case study:
  - model compression of more than factor 2 possible
  - accuracy loss is minimal (0.23%)
- can also be used for compression of model ensembles



pixy.org, CC BY-NC-ND 4.0