Learning from a teacher

Distillation

Cognitive Tunnel Effect: between conceptual domains

Coaching

Antoine Cornuéjols

AgroParisTech – INRAE UMR MIA Paris-Saclay

antoine.cornuejols@agroparistech.fr

Outline

Learning from a teacher

- **Classical** inductive learning
  - Data set
  - Prior knowledge given as a **bias**
    - E.g. prefer simple hypotheses

- Learning from a **teacher**
  - Data set
  - **Knowledge** provided by the **teacher**
    - How the teacher **processes** the queries
    - **Answers** that the learner should try to “copy”
    - **Additional** information

Outline

1. Learning from a teacher
2. Distillation
3. Cognitive tunnel effect
4. Coaching

Learning Neural Networks using “distillation”

[HINTON, Geoffrey. **Distilling the Knowledge in a Neural Network**. arXiv preprint arXiv:1503.02531, 2015.]

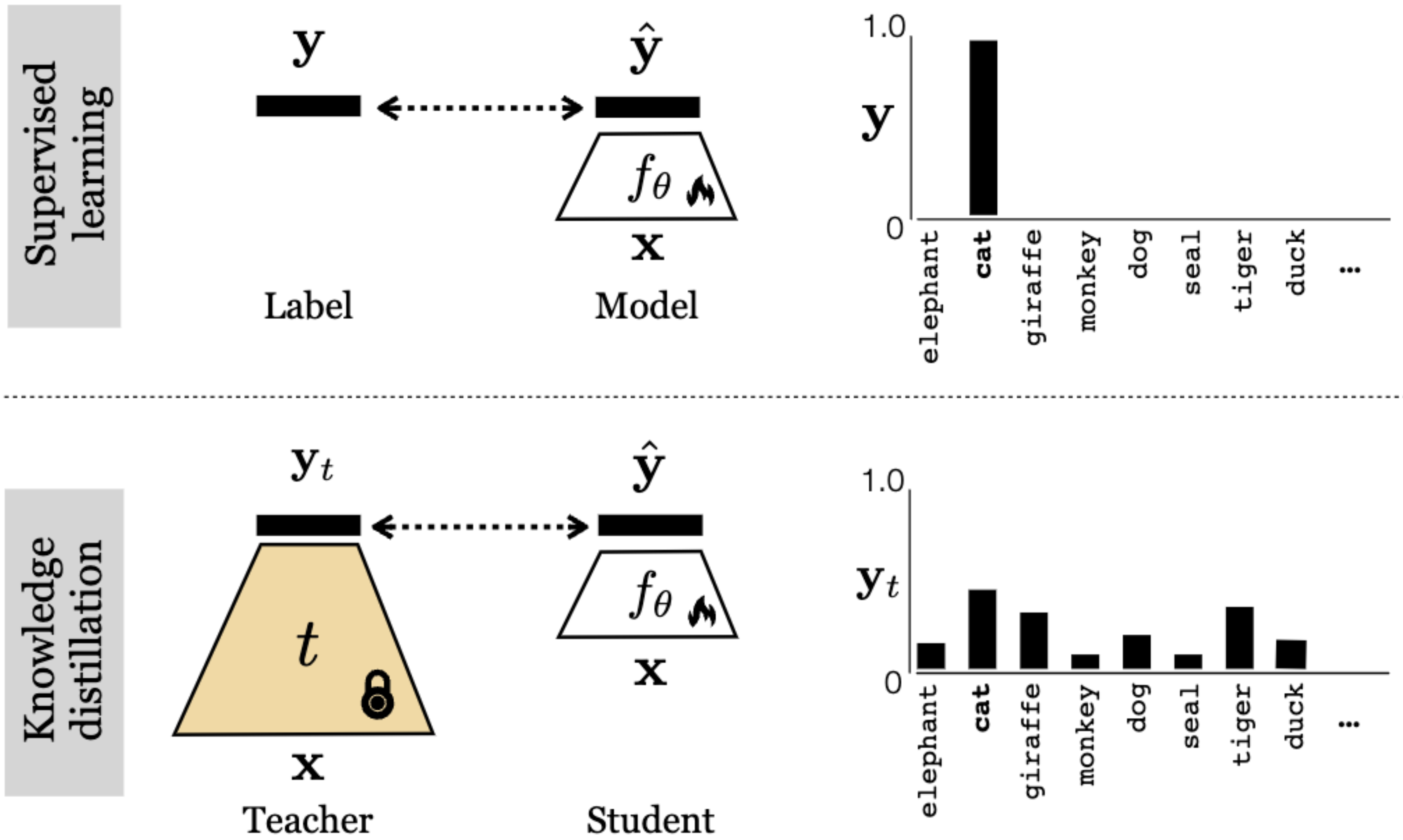
Motivation

1. We would like to deploy a classifier (NN) on a **computationally limited device** (e.g. *a smartphone*)
   – A deep NN cannot be used

2. The **learning task is difficult** and requires a large data set and a sophisticated learning method (e.g. a deep NN)

Question: can we use the learned deep NN as a **teacher** to help the **student** (i.e. the limited device) learn a simpler classifier?

Knowledge distillation: principle

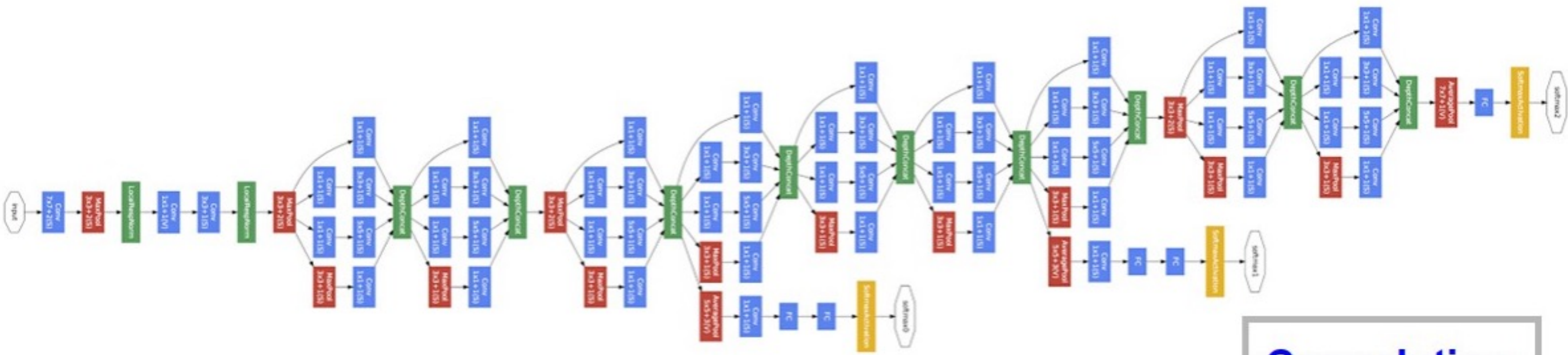


...

From [Antonio Torralba et al. **Foundations of Computer Vision**. MIT Press, 2024, p.550]

Motivation

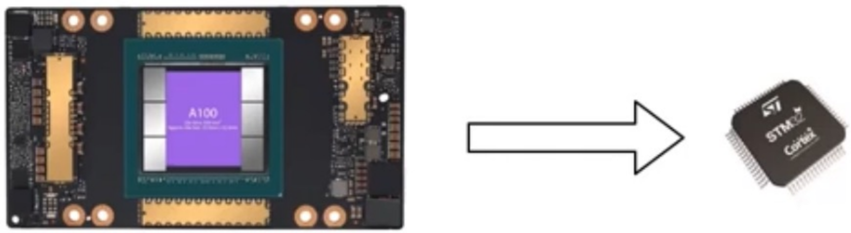
Example: A sophisticated learning technique - GoogLeNet



Quite a costly machine to train

AND to use for prediction

Motivation

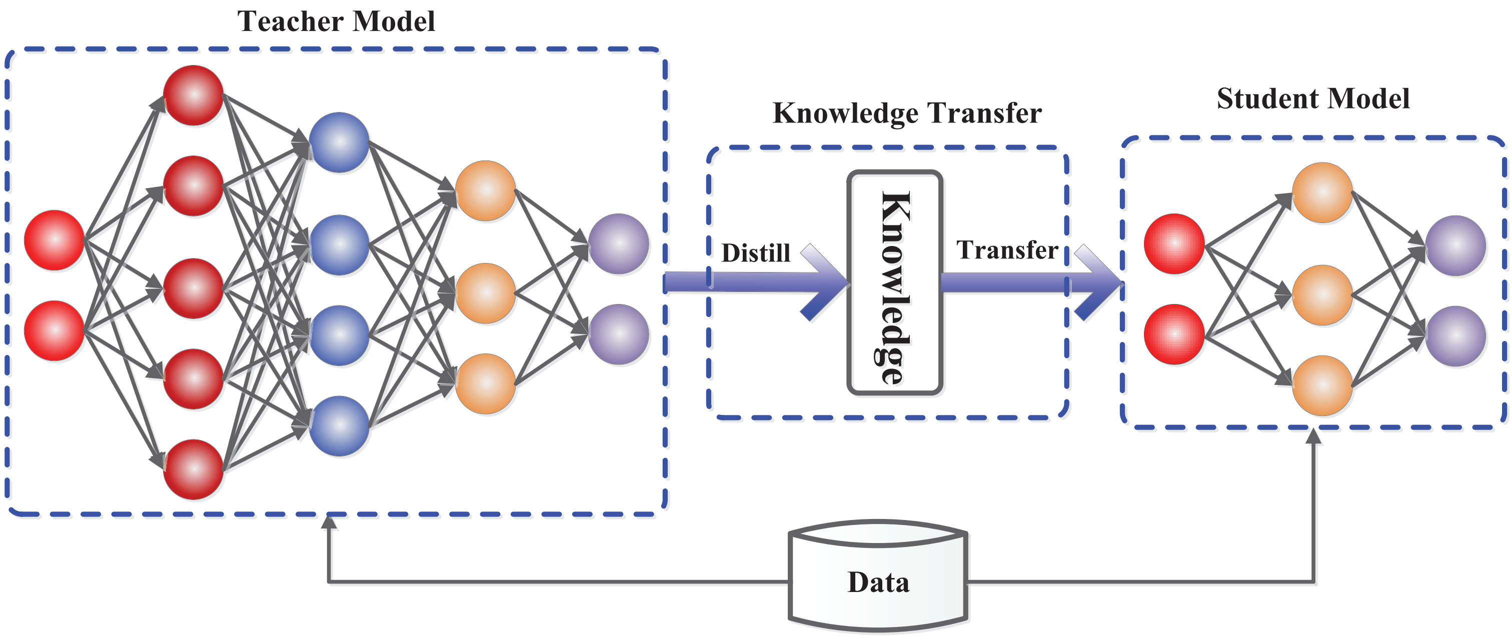


	Cloud AI	Tiny AI
Computation (fp32)	19.5 TFLOPS	MFLOPs
Memory	80GB	256kB
Neural Network	ResNet ViT-Large ...	MCUNet MobileNetV2-Tiny ...

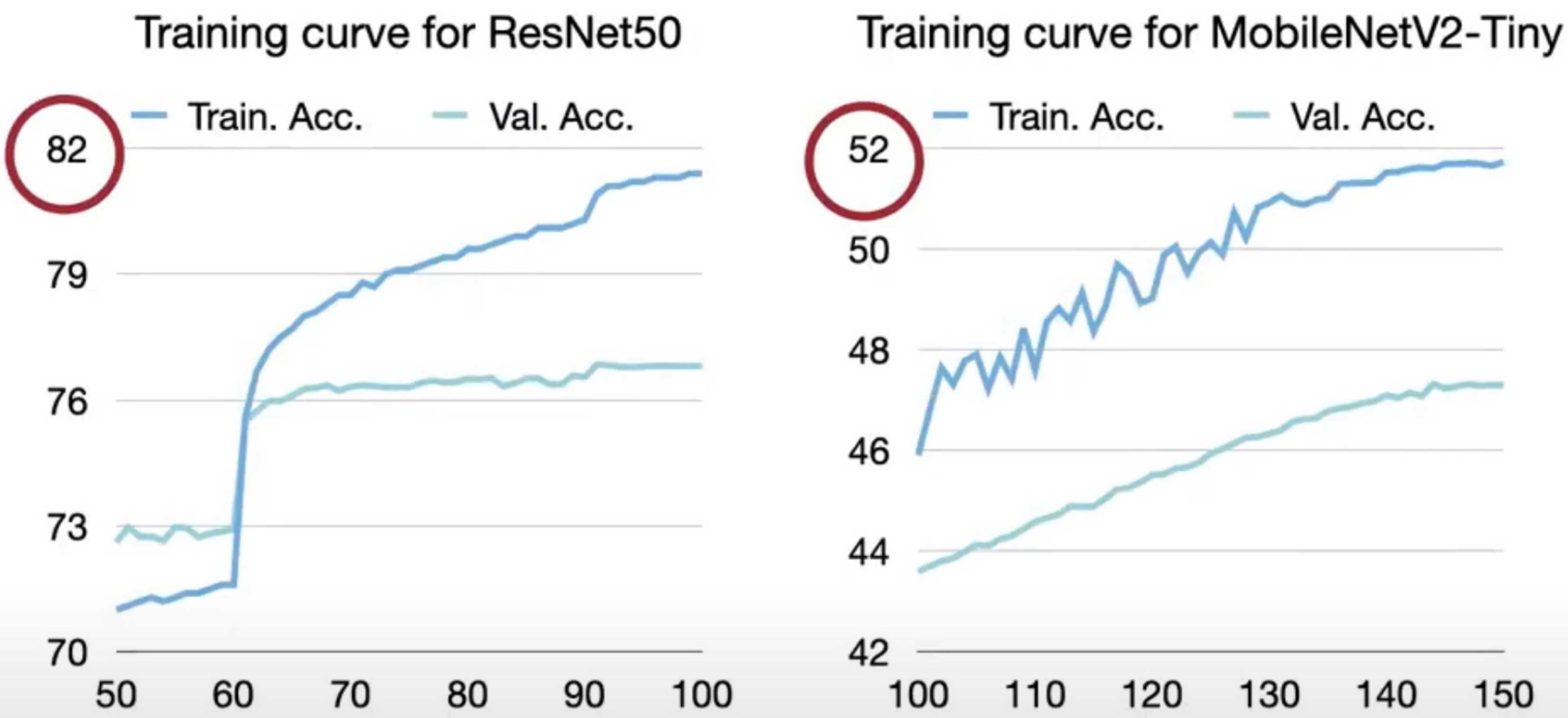
- Neural network must be **tiny** to run efficiently on tiny edge devices.

...

Knowledge distillation



...



Question: Can we help the training of tiny models with large models?

...

Learning techniques for "distillation"

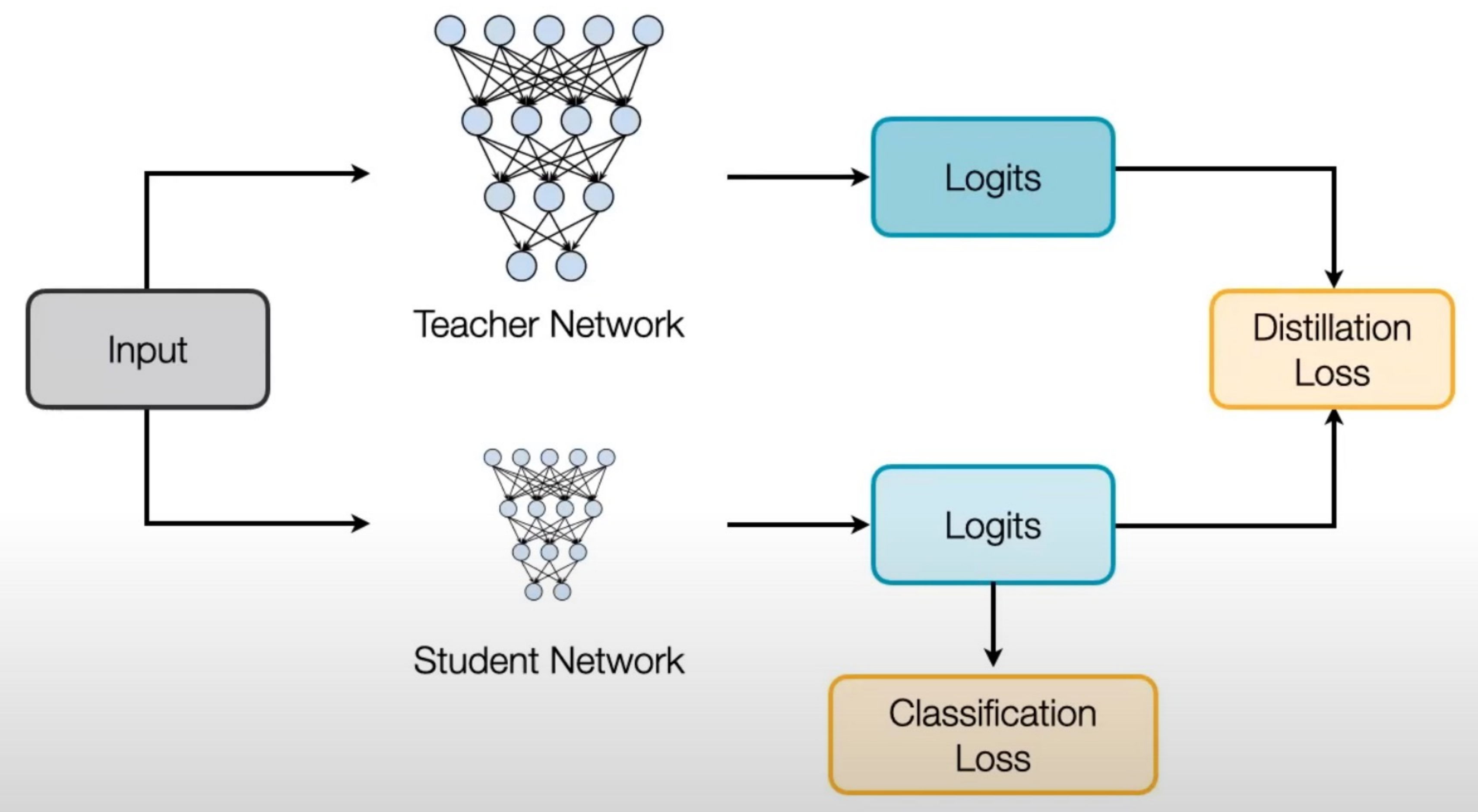
1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**

Not exhaustive. For a survey, see:

[GOU, Jianping, YU, Baosheng, MAYBANK, Stephen J., et al. **Knowledge distillation: A survey**. International Journal of Computer Vision, 2021, vol. 129, no 6, p. 1789-1819.]

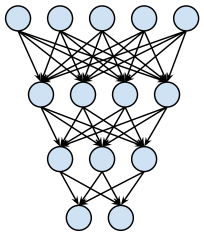
Learning techniques for “distillation”

1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**


Input
Teacher Network
Student Network
Logits
Logits
Distillation Loss
Classification Loss

…

Matching the **targets**

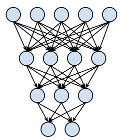


Teacher Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$\frac{\exp(5)}{\exp(5)+\exp(1)}$$

$$\frac{\exp(1)}{\exp(5)+\exp(1)}$$



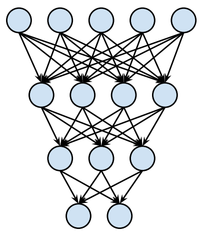
Student Network

	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

...

Matching the **targets**



Teacher Network

$$\frac{\exp(5/1)}{\exp(5/1) + \exp(1/1)}$$

	Logits	Probabilities (T=1)	Probabilities (T=10)
Cat	5	0.982	0.599
Dog	1	0.017	0.401

$$\frac{\exp(5/10)}{\exp(5/10) + \exp(1/10)}$$

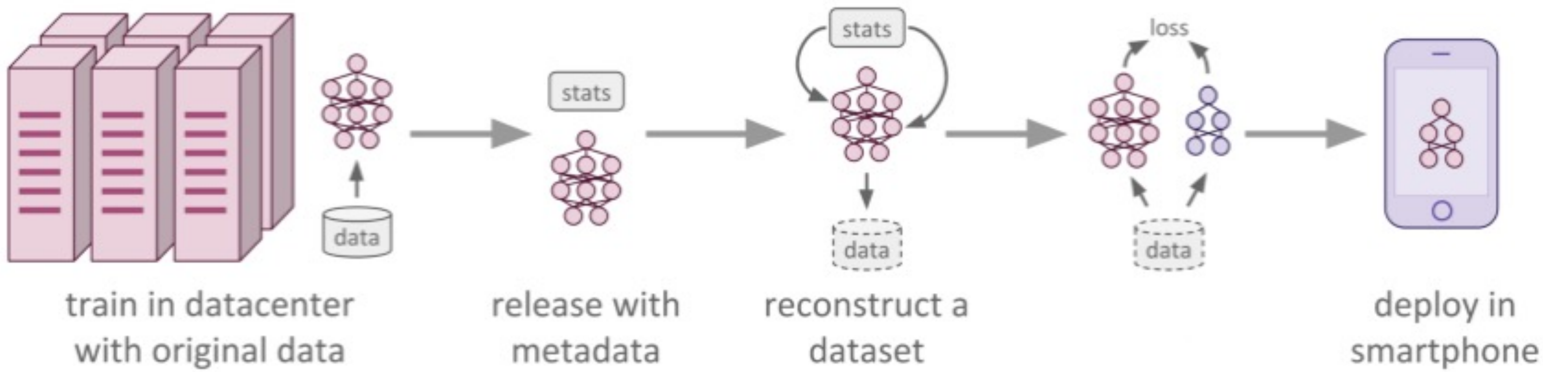
A larger temperature smooths the output probability distribution.

...

Changing the **target**

1. Use the sophisticated learning method (teacher) to learn to predict the target classes with a **membership measure**

2. Ask the student to *learn to predict the **membership measure*** computed by the teacher instead of the hard classes (on the training set)



Changing the target

1. The teacher uses a softmax function for the values of its output

$$q_i = \frac{e^{(z_i/T)}}{\sum_{j \in \text{classes}} e^{(z_j/T)}}$$

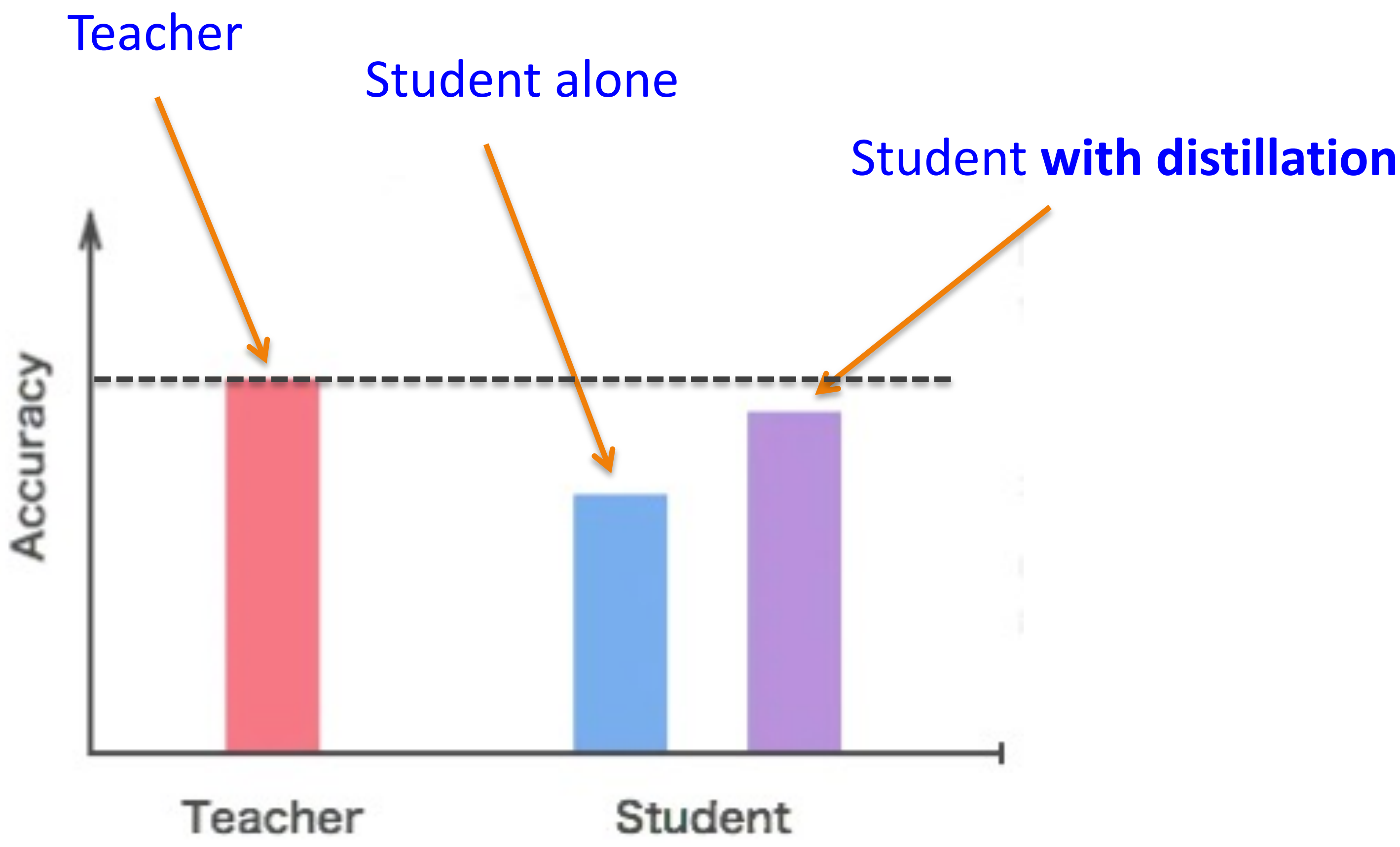
   T is the temperature (the highest T, the less different are the outputs)

2. The student *learns to predict the membership measure* first with T high, and then, progressively, with T decreasing to 1.

When the soft targets have high entropy, they **provide much more information per training case** than hard targets and **much less variance in the gradient** between training cases, so the small model can often be trained on much less data than the original cumbersome model while using a much higher learning rate.

Changing the target

...



Learning techniques for “distillation”

1. Matching the **targets**

2. Matching intermediate **weights**

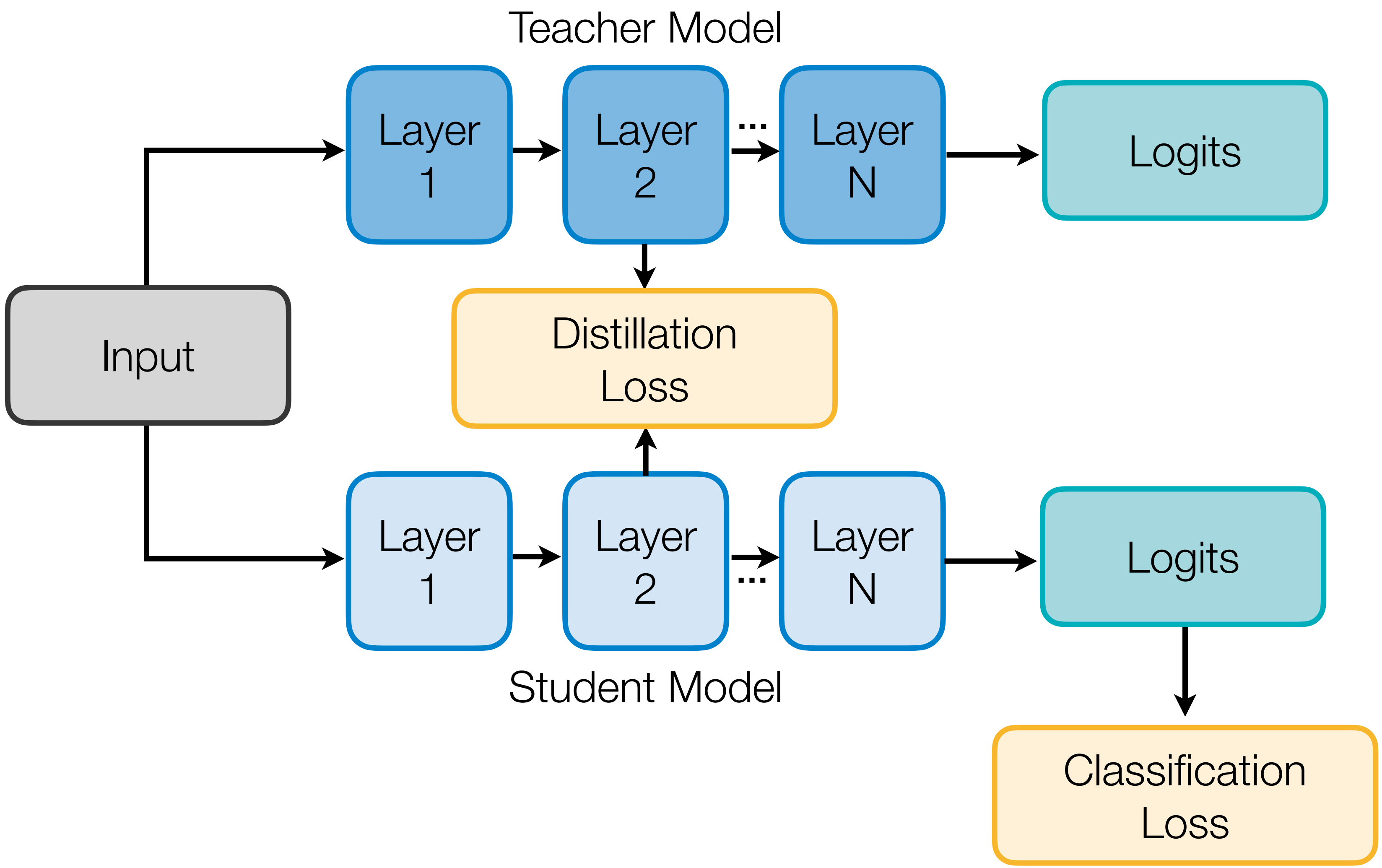
3. Matching intermediate **features**

4. Matching **gradients**

5. Gradually changing the **inputs**

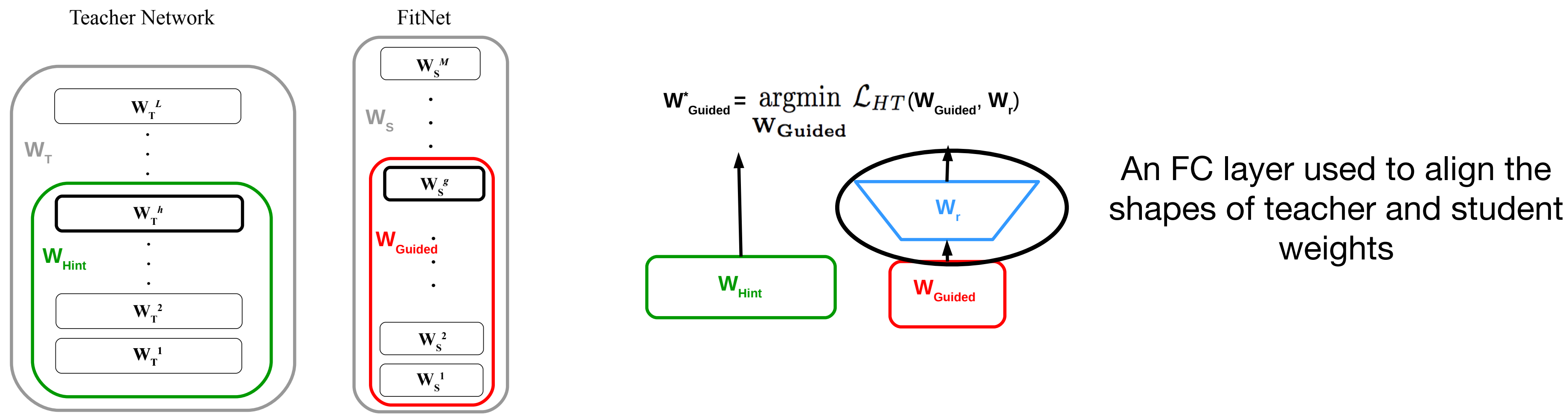
6. Gradually changing the **learning task**

Matching intermediate **weights**



...

Matching intermediate **weights**



(a) Teacher and Student Networks (b) Hints Training

The cross-entropy loss (classification)

\+ a L2 loss between **teacher** weights and **student** weights

FitNets: Hints for Thin Deep Nets [Romero *et al.*, ICLR 2015]

Learning techniques for “distillation”

1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**

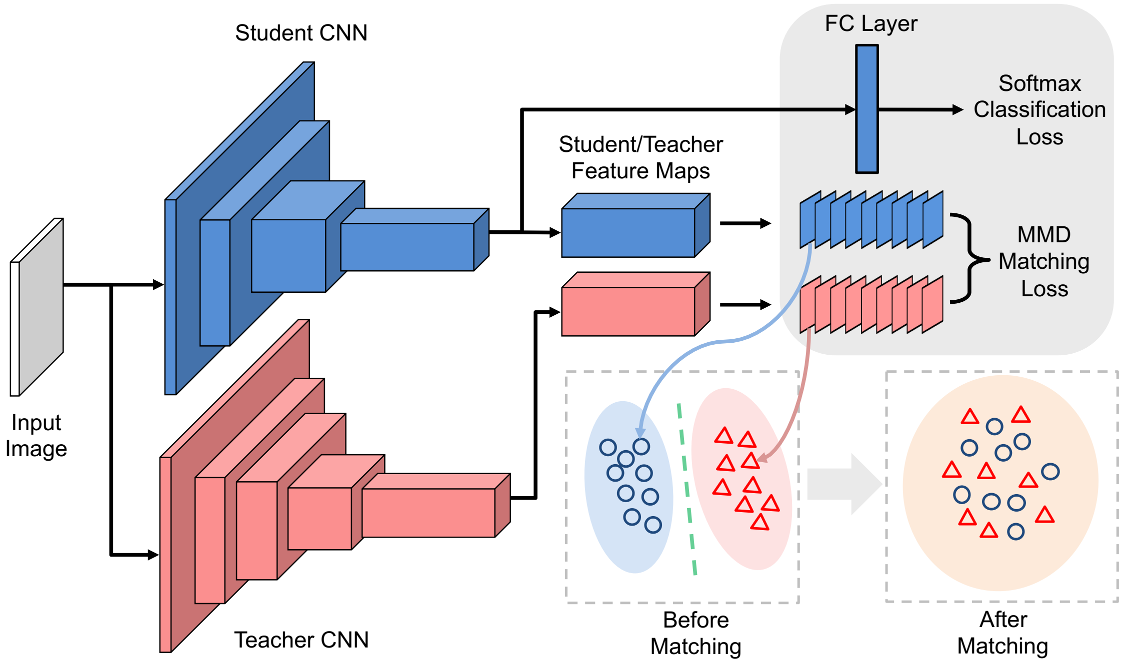
Matching intermediate **features**

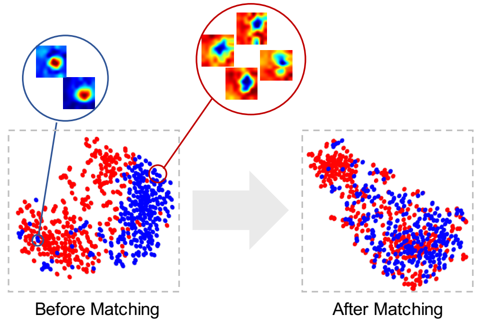
- Motivation
  - Each neuron essentially extracts a **certain pattern** related to the task at hand from raw input.
    - If a neuron is activated in certain regions, that implies these regions share some common properties that may relate to the task. It **provides a kind of explanation** to the final prediction of the teacher model.
  - Therefore, try to **align** the distribution of neuron **selectivity pattern** between student models and teacher models.

(a) Monkey (b) Magnetic Hill

Figure 2. Neuron activation heat map of two selected images.

Matching intermediate **features**





t-SNE [31] visualization shows that NST Transfer reduces the distance between teacher and student activations distribution.

The architecture for the **Neuron Selectivity Transfer**: the student network is not only trained from ground-truth labels, but it also mimics the distribution of the activations from **intermediate layers** in the teacher network (by minimizing the Maximum Mean Discrepancy).

Each dot or triangle in the figure denotes its corresponding activation map of a filter.

Learning techniques for “distillation”

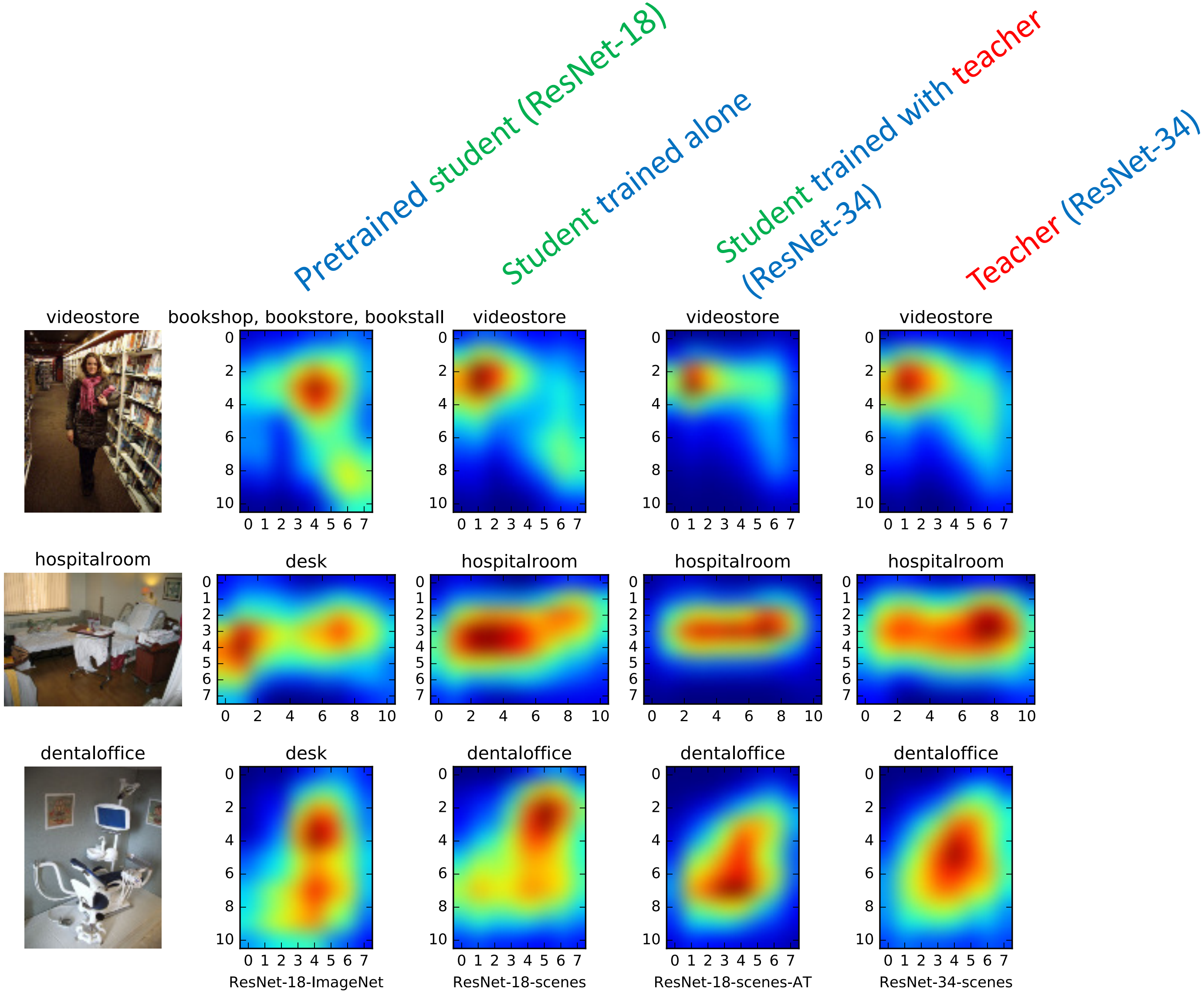
1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**

Matching **gradients**

- Similar motivation

Figure 1: **(a)** An input image and a corresponding spatial attention map of a convolutional network that shows where the network focuses in order to classify the given image. Surely, this type of map must contain valuable information about the network. The question that we pose in this paper is the following: can we use knowledge of this type to improve the training of CNN models ? **(b)** Schematic representation of attention transfer: a student CNN is trained so as, not only to make good predictions, but to also have similar spatial attention maps to those of an already trained teacher CNN.

ZAGORUYKO, Sergey et KOMODAKIS, Nikos. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.



ZAGORUYKO, Sergey et KOMODAKIS, Nikos. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Learning techniques for "distillation"

1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**

Changing the **inputs**

- Idea: friendly training vs. adversary learning
  - Modifies the inputs so as **to facilitate** the training
- **Modifies** the descriptions of the **examples**
  - According to the current training stage $\tilde{x}_i = x_i + \delta_i$
  - So as to minimize: $$L(\mathcal{B}, w) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \ell\big(f(\tilde{x}_i, w), y_i\big)$$

Marullo, S., Tiezzi, M., Gori, M., & Melacci, S. (2021). **Being Friends Instead of Adversaries: Deep Networks Learn from Data Simplified by Other Networks**. *arXiv preprint arXiv:2112.09968.*

Neural Friendly Training

- But the modifications are **independently** applied to all training examples
- We would rather like global deformations that help to learn the decision function

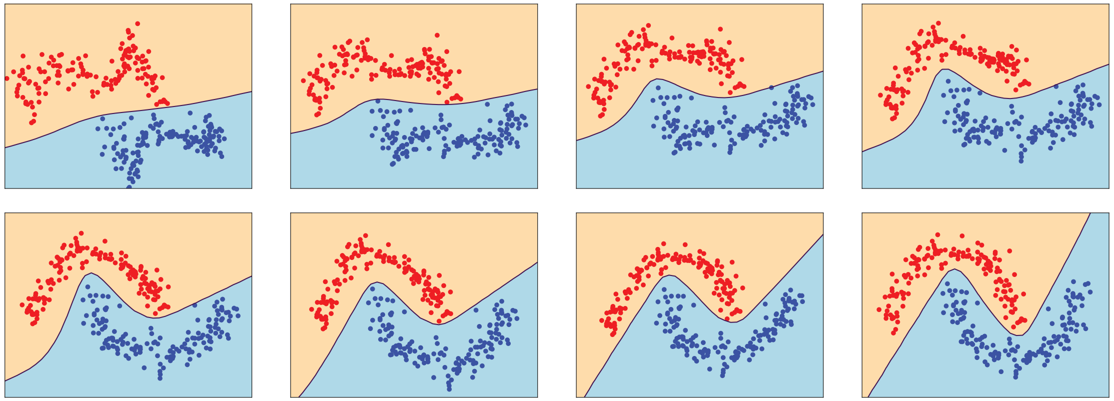
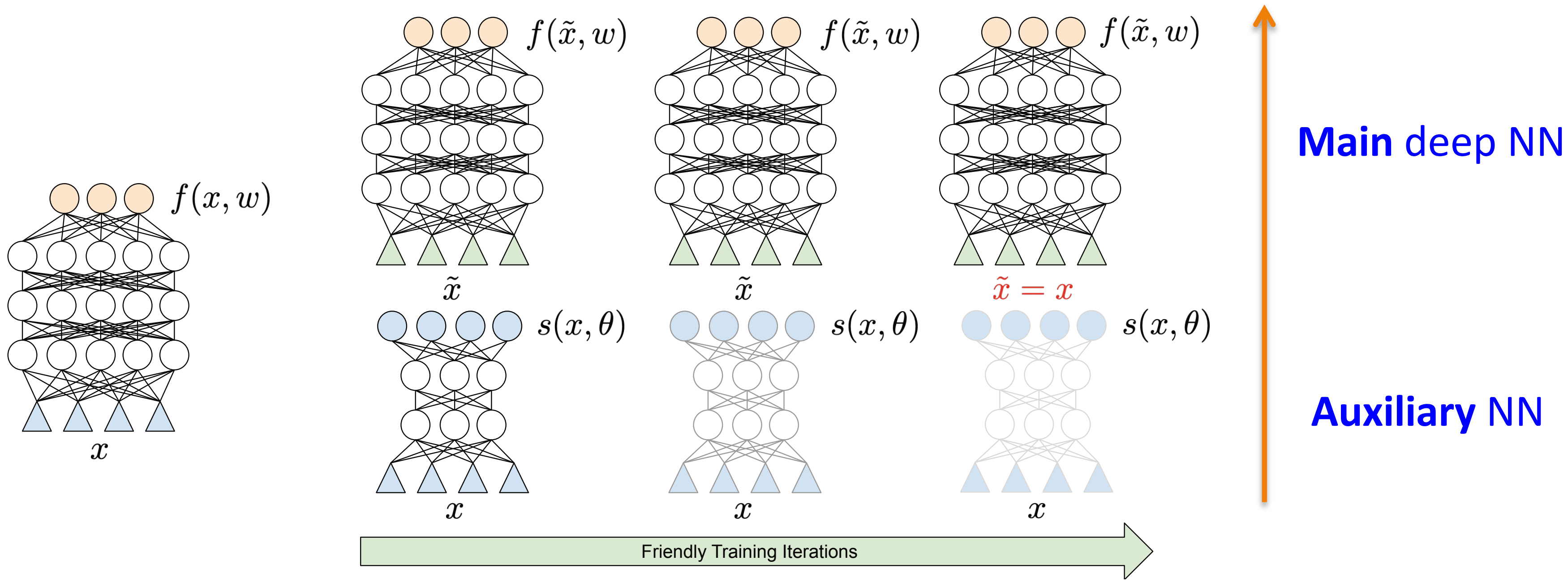


Figure 1: Left-to-right, top-to-bottom: evolution of the decision boundary developed by a single hidden layer classifier (5 neurons) in the 2-moon dataset, in Neural Friendly Training. Each plot is about a different training iteration (γ); in the last plot data are not transformed anymore.

$$\tilde{x}_i = s(x_i, \theta)$$

Neural Friendly Training



$$L(\mathcal{B}, w, \theta) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \Bigg(\ell\big(f(\underbrace{s(x_i, \theta)}_{\tilde{x}_i}, w), y_i\big) + \eta \Big\| \underbrace{s(x_i, \theta) - x_i}_{\delta_i} \Big\|^2 \Bigg),$$

...

Neural Friendly Training

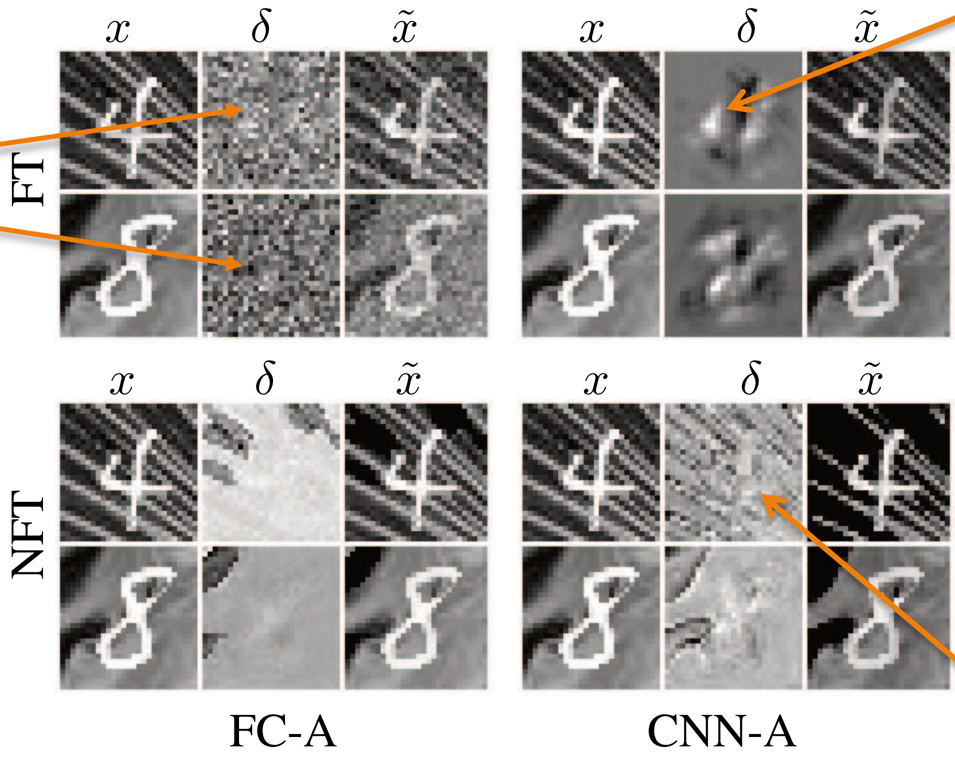
FC-A: **F**ully **C**onnected MLP

CNN-A: Convolutional **NN**

Structured perturbations with CNNs only, **emphasizing the digit areas**

Really **poor**

Globally more **satisfying**



FT: **F**riendly **T**raining

Independent transformation for each example

NFT: **N**eural **F**riendly **T**raining

Using an **auxiliary** NN

Perturbations **removing distracting cues**

Figure 4: MNIST-BACK-IMAGE. Original data x , perturbation δ (normalized) and resulting "simplified" images $\tilde{x}$ for FC-A and CNN-A at the end of the 1st epoch. Some simplifications are hardly distinguishable. Top: FT. Bottom: NFT.

...

Learning techniques for “distillation”

1. Gradually changing the **targets**

2. Gradually changing the **inputs**

3. Gradually changing the **learning task**

Learning techniques for “distillation”

1. Matching the **targets**
2. Matching intermediate **weights**
3. Matching intermediate **features**
4. Matching **gradients**
5. Gradually changing the **inputs**
6. Gradually changing the **learning task**

Changing the learning task

- The **classical** distillation scenario (adapted)

Stationary!

$$\mathcal{L}_{KD} = (1-\alpha)\underbrace{H(y, q_s(\theta))} + \alpha T^2 \underbrace{H(p_t, q_s(\theta))}$$

Classical cross-entropy between **output** and **target** values

Cross-entropy between **teacher** and **student's** outputs

Changing the learning task

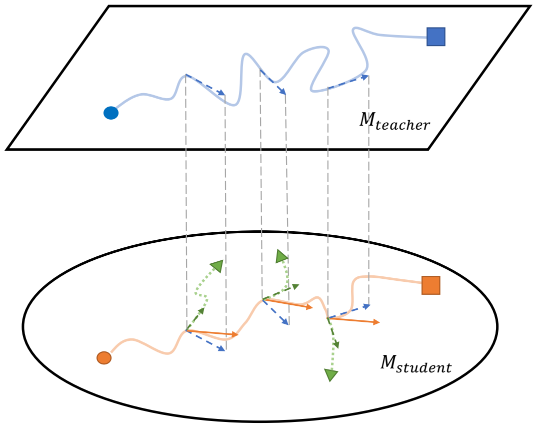
- Idea: train the student network through a sequence of **intermediate** learning **tasks**.
- Question: **how to choose** the intermediate learning tasks?
  1. They should be **easily achievable** by the student
  2. Consequence: the **teacher should be aware** of the student's progress

Changing the learning task

- Idea: train the student network through a sequence of **intermediate** learning **tasks**.
- Question: **how to choose** the intermediate learning tasks?
  1. They should be **easily achievable** by the student
  2. Consequence: the **teacher should be aware** of the student's progress
- Co-evolution between student and teacher
  1. The teacher converges toward the **goal**, but stay close to the **learner**

$$\theta_t^{m+1} = \min_{\theta_t} H(y, p_{\theta_t}) \quad \text{s.t. } D_{\mathrm{KL}}(q_{\theta_s}^m, p_{\theta_t}) \leq \epsilon$$

$$\hat{\mathcal{L}}_{\theta_t} = (1-\lambda)H(y, p_{\theta_t}) + \lambda H(q_{\theta_s}, p_{\theta_t})$$



  2. The student follows the **teacher** at each step

$$\theta_s^{m+1} = \theta_s^m - \eta_s \nabla \mathcal{L}_s(\theta_s, p_{\theta_t^{m+1}}), \quad \mathcal{L}_s(\theta_s) = H(p_{\theta_t}, q_{\theta_s})$$

Changing the learning task

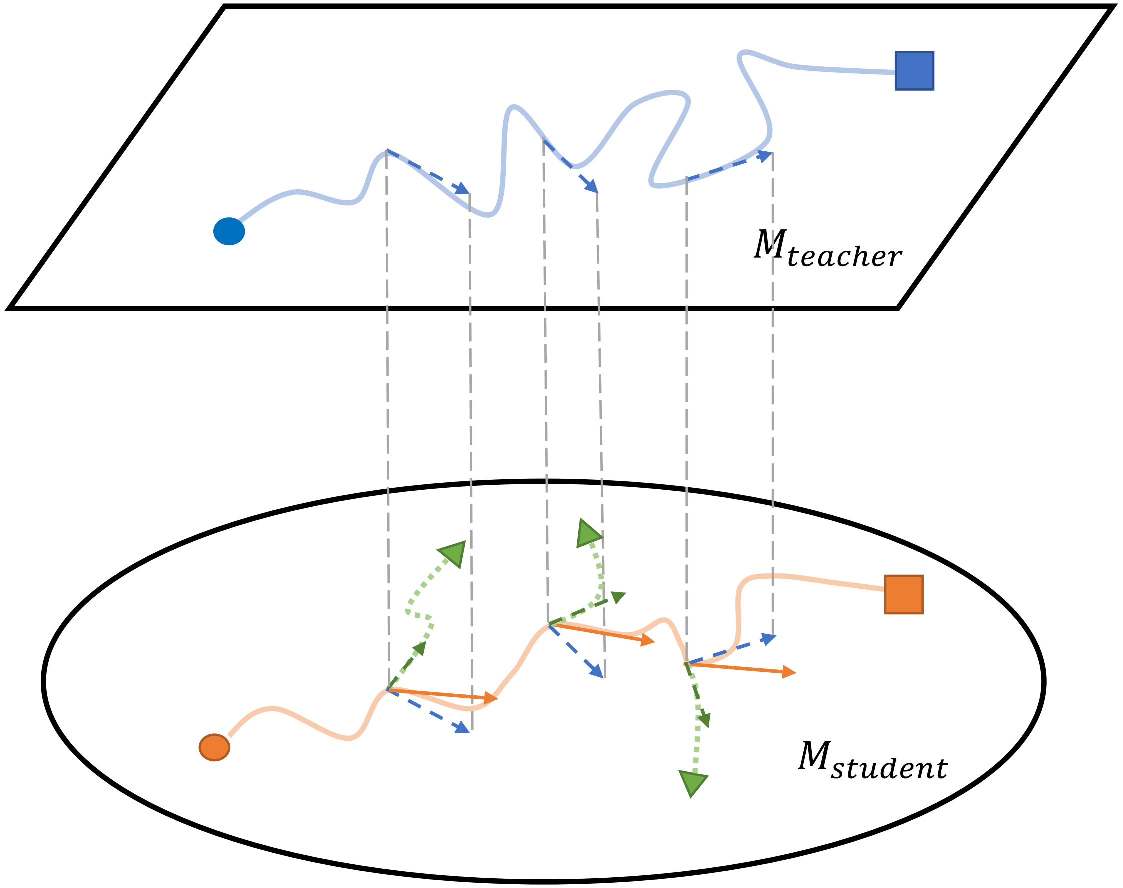


Fig. 1: $\mathcal{M}_{teacher}$ and $\mathcal{M}_{student}$ refer to the output manifolds of student model and teacher model. The lines between circles (●,●) to squares (■,■) imply the learning trajectories in the distribution level. The intuition of ProKT is to avoid bad local optimas (triangles (▲)) by conducting supervision signal projection.

...

Changing the learning task

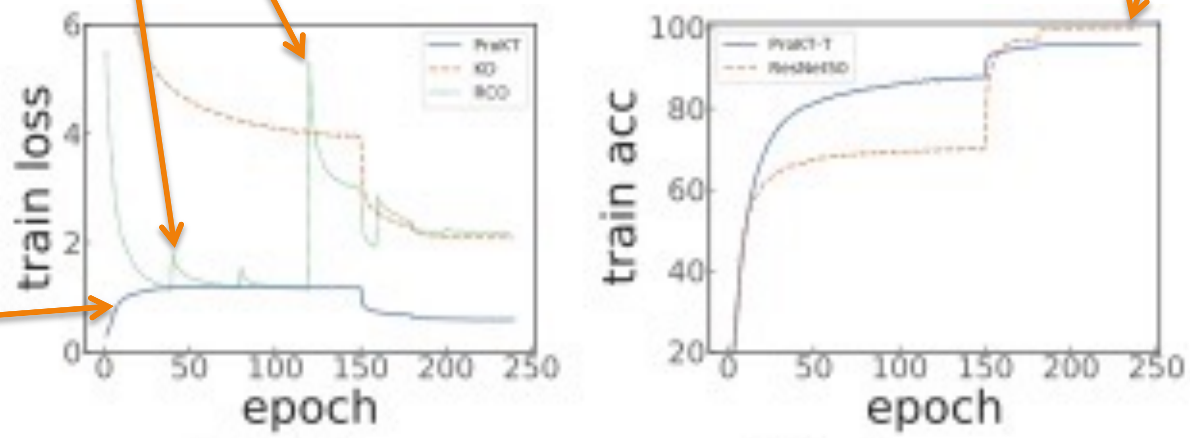
KD : classical **K**nowledge **D**istillation

RCO : use intermediate models obtained during the teacher's training process

ProKT : their method

Lower performance of the teacher, but better student in the end

The divergence between teacher and student in ProKT is **smooth** and **well bounded**



(a) Train loss of **student** (b) Train accuracy of **teacher**

Shi, W., Song, Y., Zhou, H., Li, B., & Li, L. (2021, September). **Follow your path: a progressive method for knowledge distillation**. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 596-611). Springer.

Changing the learning task

KD : classical **K**nowledge **D**istillation

RCO : use intermediate models obtained during the teacher's training process

ProKT : their method where the teacher stays close to the student

Using Kullback-Leibler (KD)loss

Without distillation

With distillation

Teacher Student	vgg13 MobileNetV2	ResNet50 MobileNetV2	ResNet50 vgg8	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.6	64.6	70.36	70.5	71.82	70.5
KD*	67.37	67.35	73.81	74.07	74.45	74.83
RCO	68.42	68.95	73.85	75.62	**76.26**	75.53
ProKT	**68.79**	**69.32**	**73.88**	**75.79**	75.59	**76.02**
CRD	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD	**69.94**	69.54	74.58	75.12	76.05	76.27
CRD+ProKT	69.59	**69.93**	**75.14**	**76.0**	**76.86**	**76.76**

Using Constrastive Representation Distillation (CRD) loss

Shi, W., Song, Y., Zhou, H., Li, B., & Li, L. (2021, September). **Follow your path: a progressive method for knowledge distillation**. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 596-611). Springer.

Lessons

- **Careful** distillation is useful

- Points to the idea of **curriculum** learning

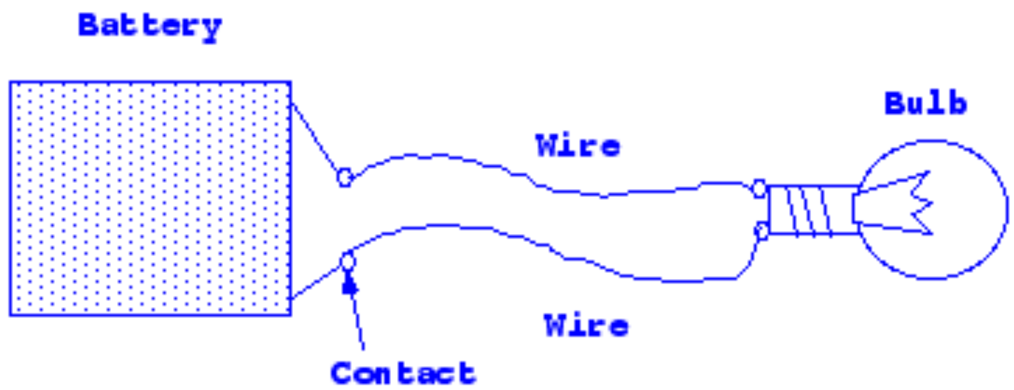
Outline

1. Learning from a teacher
2. Distillation
3. Cognitive tunnel effect
4. Coaching

Cognitive tunnel effect

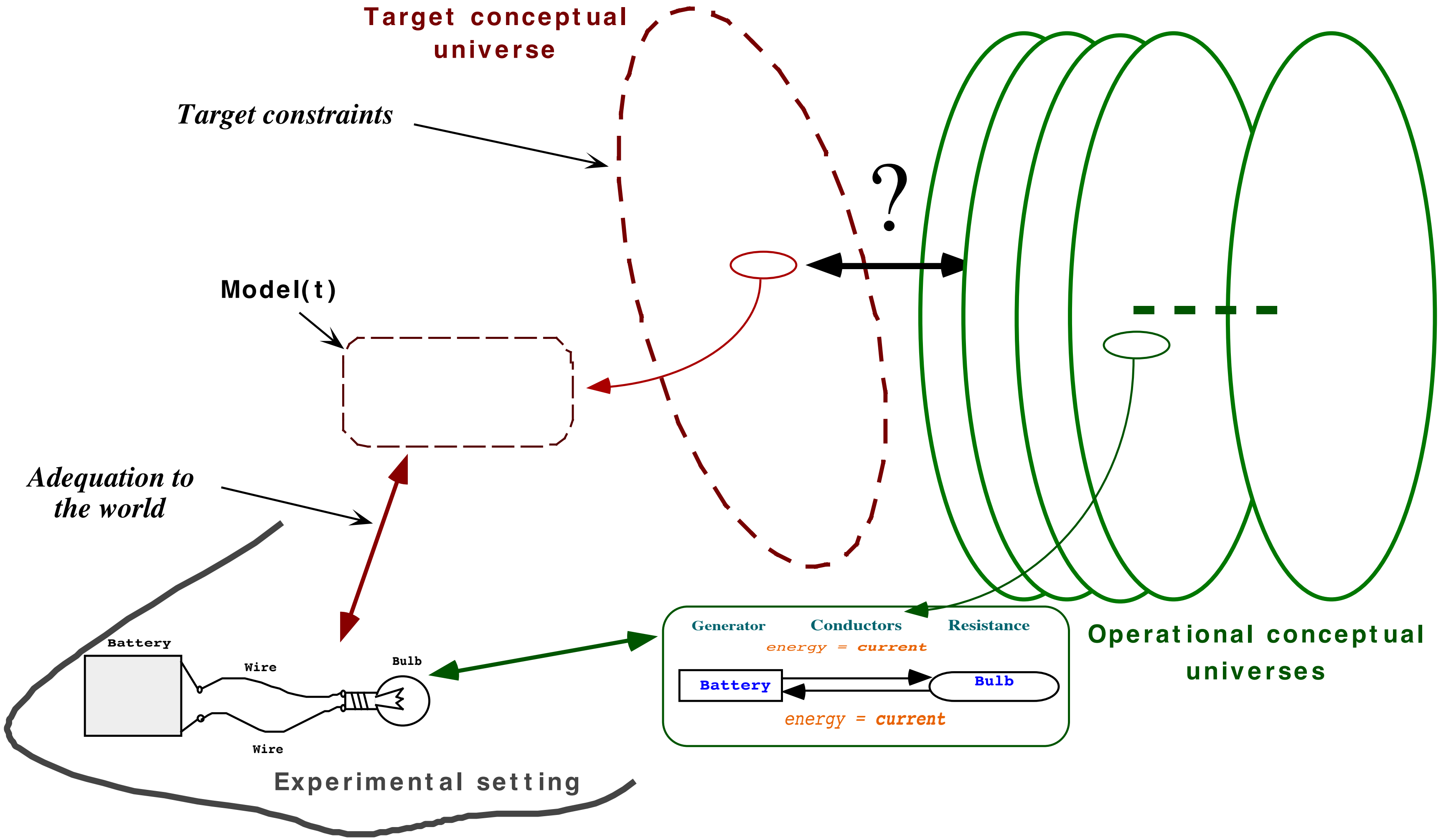
[A. Cornuéjols, A. Tiberghien, G. Collet. *Tunnel Effects in Cognition: A new Mechanism for Scientific Discovery and Education*. Arxiv-1707.04903- Tue, 18 Jul 2017 00:00:00 GMT]

Experimental
setting



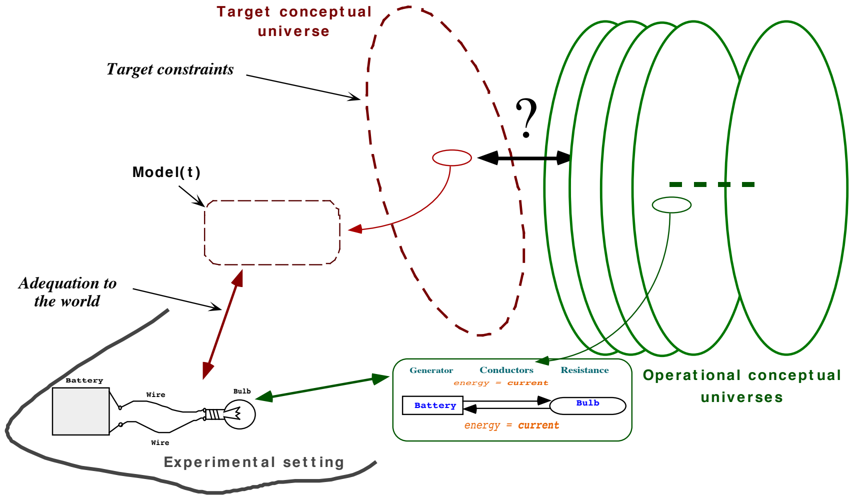
Interpret this experiment in terms of energy transfers

Cognitive tunnel effect

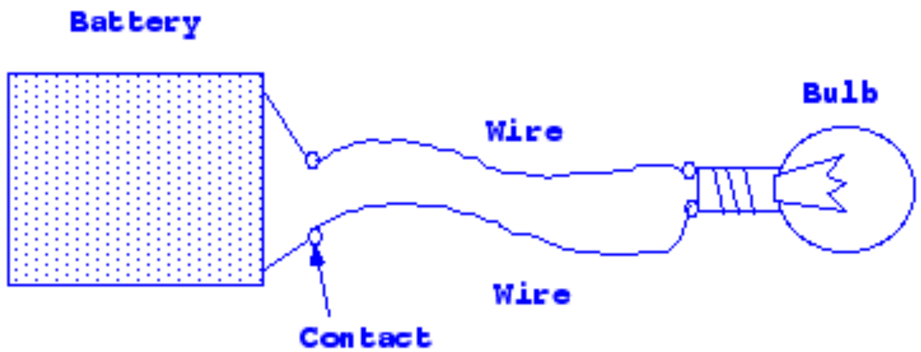


...

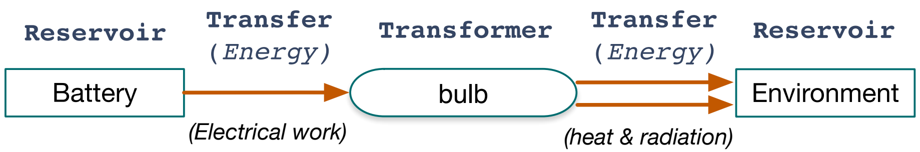
Cognitive tunnel effect



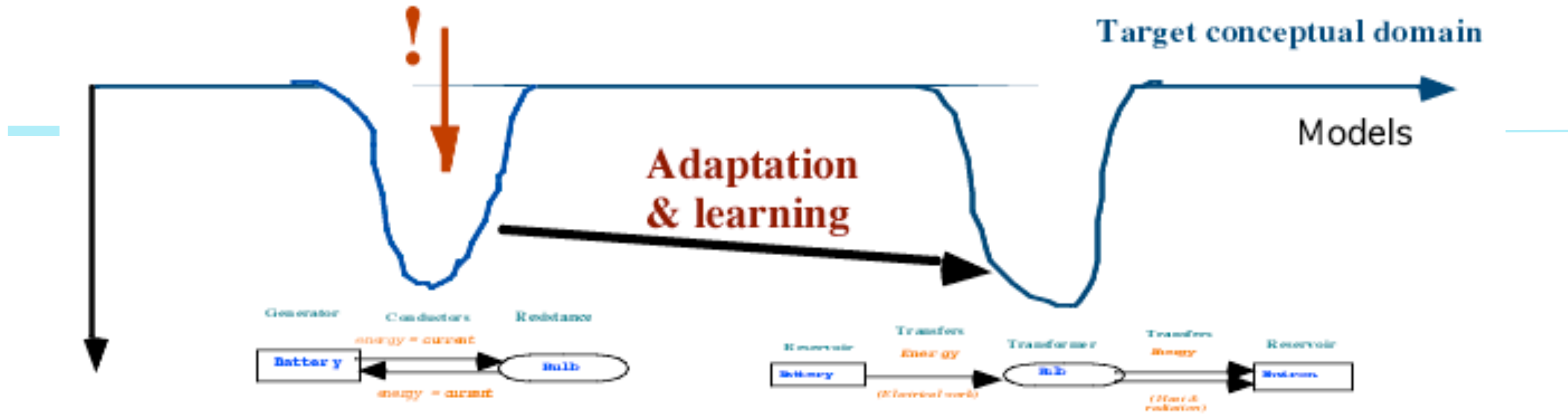
Experimental setting

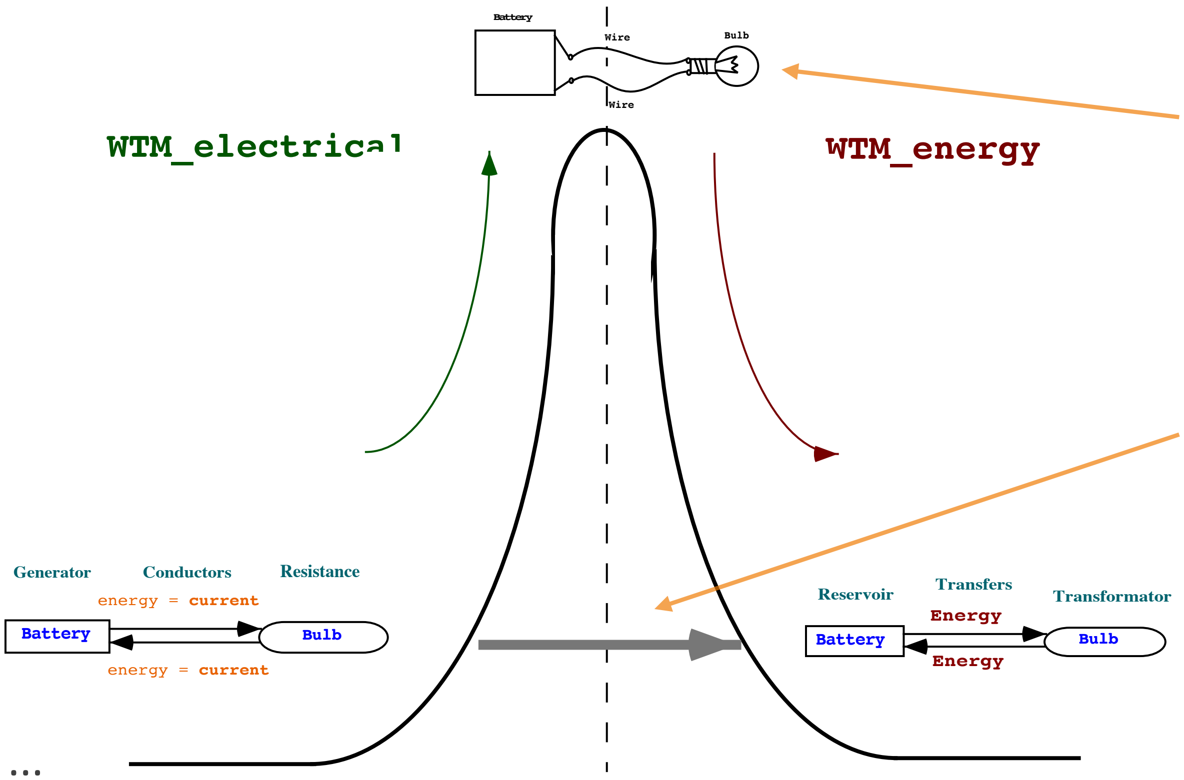


Conceptual interpretation in terms of **energy chain**



...





Students do not come back to the “naked” representation to interpret the setting in the new domain.

They dig a cognitive tunnel to the new conceptual domain **smuggling in** interpretations from the source domain (e.g. the arrows), **then** trying to make it work in the new domain.

Newton’s luggage

[Loup Verlet. La malle de Newton. Gallimard, NRF, 1993]

- How did Newton arrive to **his theory of gravitation**?
- What were the **sources** of his thoughts?
  - **Alchemy** (among other things), but …
- What were the **questions of the time**?
  - How **transmutation of bread** into the corpse of Jesus Christ can arise at the “same time” in all churches on Sunday services?
    - In **Britain**, simultaneously ---> Action **at distance**
    - In **continental Europe**, depending on signals arriving at the church ---> Action in **need of a medium**

Conclusion

- Co-learning
  - Assumption: there are two (or more) complementary views (description spaces)
- Boosting
  - Assumption: **changing the input distribution** allows learning a useful change of representation
- Blending
  - Assumption: **two** frames of reference can be merged to bring complementary information
- Cognitive tunnel effect
  - Assumption: a **single** representation can be interpreted within **two** worlds.
  - And the resulting cognitive obstacles can lead to progresses in building a conceptual perspective on the world.