Incremental Learning in Image Classification

Fabio Cermelli (fabio.cermelli@polito.it)

MLDL 2020

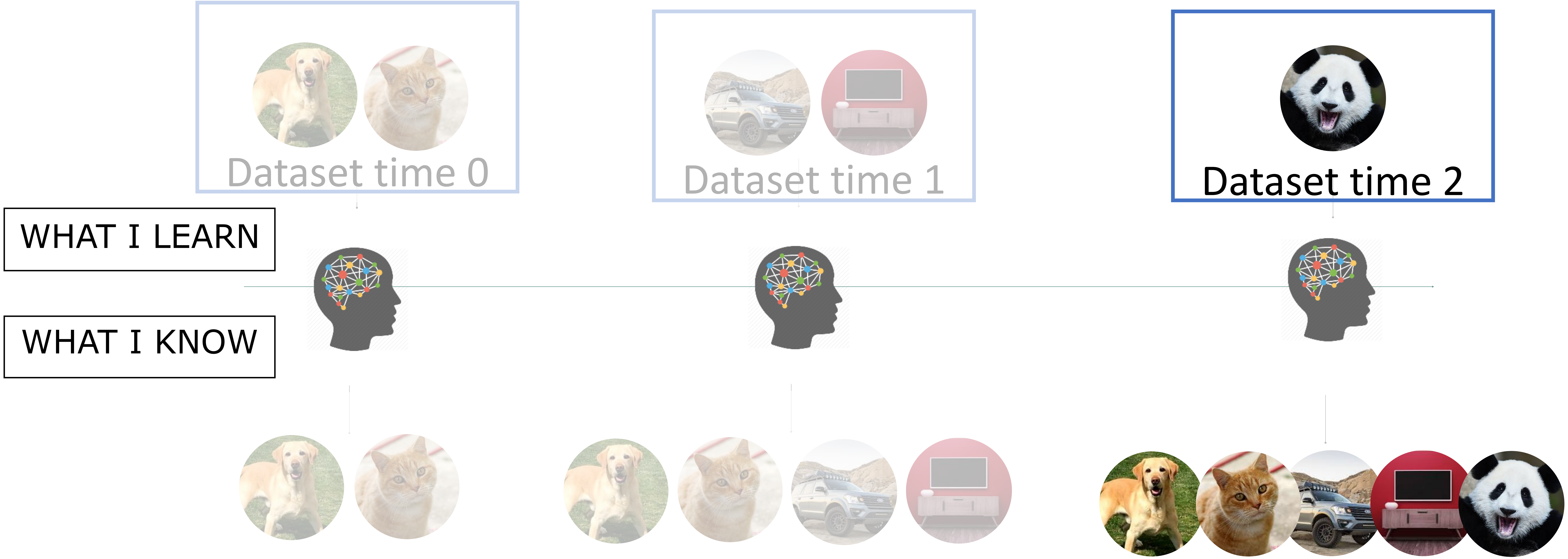
Incremental Class Learning

Some context of the project

Goal

Incremental Class Learning

Add new classes to an already trained model.



Goal

Incremental Class Learning

Add new classes to an already trained model.

Easy?

Training again the network from scratch with all the classes.

Expensive in terms of time and money!

Data from previous classes may be no more available.

Goal

Incremental Class Learning

Add new classes to an already trained model.

A cheaper solution!
Update the network parameters only with new data.

Efficient in terms of time and money!
Data of previous classes are not used.
But.....

Catastrophic Forgetting

Incremental Class Learning

When a neural network tries to extend its knowledge by learning new classes, it forgets the previously learned ones.

Robert French. Catastrophic forgetting in connectionist networks.
Trends in cognitive sciences, 3 (1999),

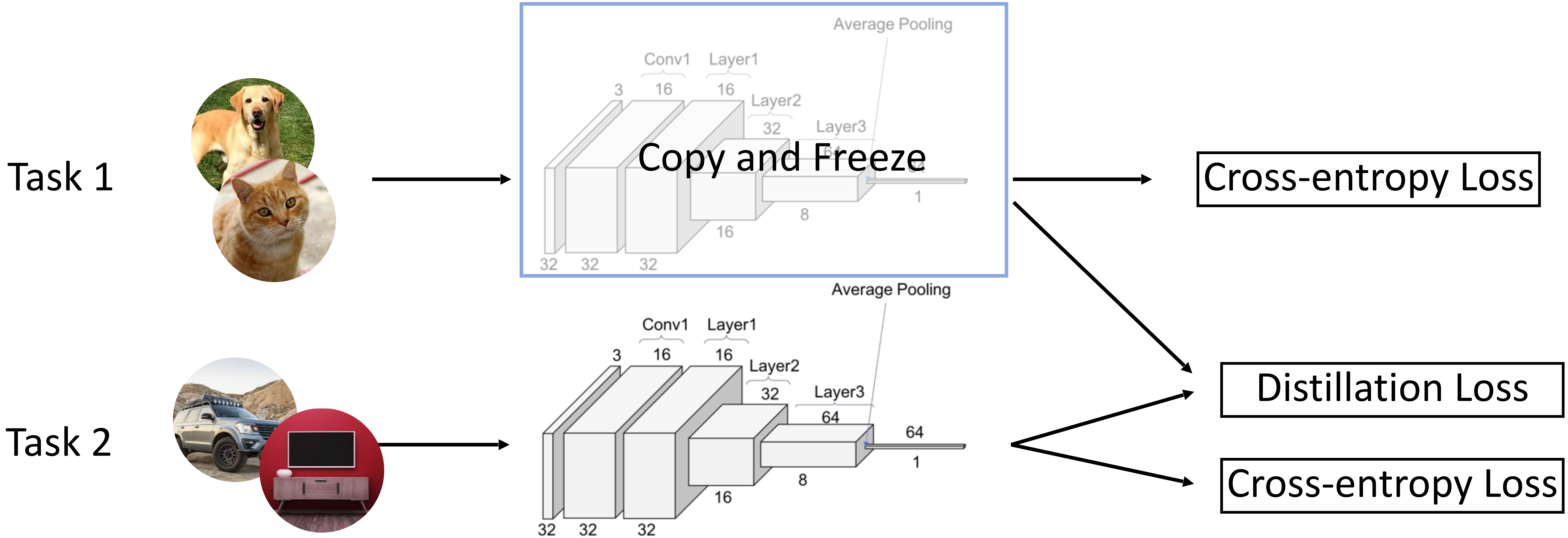
Catastrophic Forgetting

Incremental Class Learning

State of the art

Some recent papers that tackle effectively Incremental Class Learning

Learning Without Forgetting



Z. Li and D. Hoiem. Learning without forgetting. In IEEE T-PAMI, 40(12):2935–2947, 2017

Learning Without Forgetting

Knowledge Distillation Loss

$$\ell_{kd}^{\theta^t}(x) = -\sum_{c \in K^{t-1}} p_{\theta^{t-1}}^{c} \log p_{\theta^t}^{c}$$

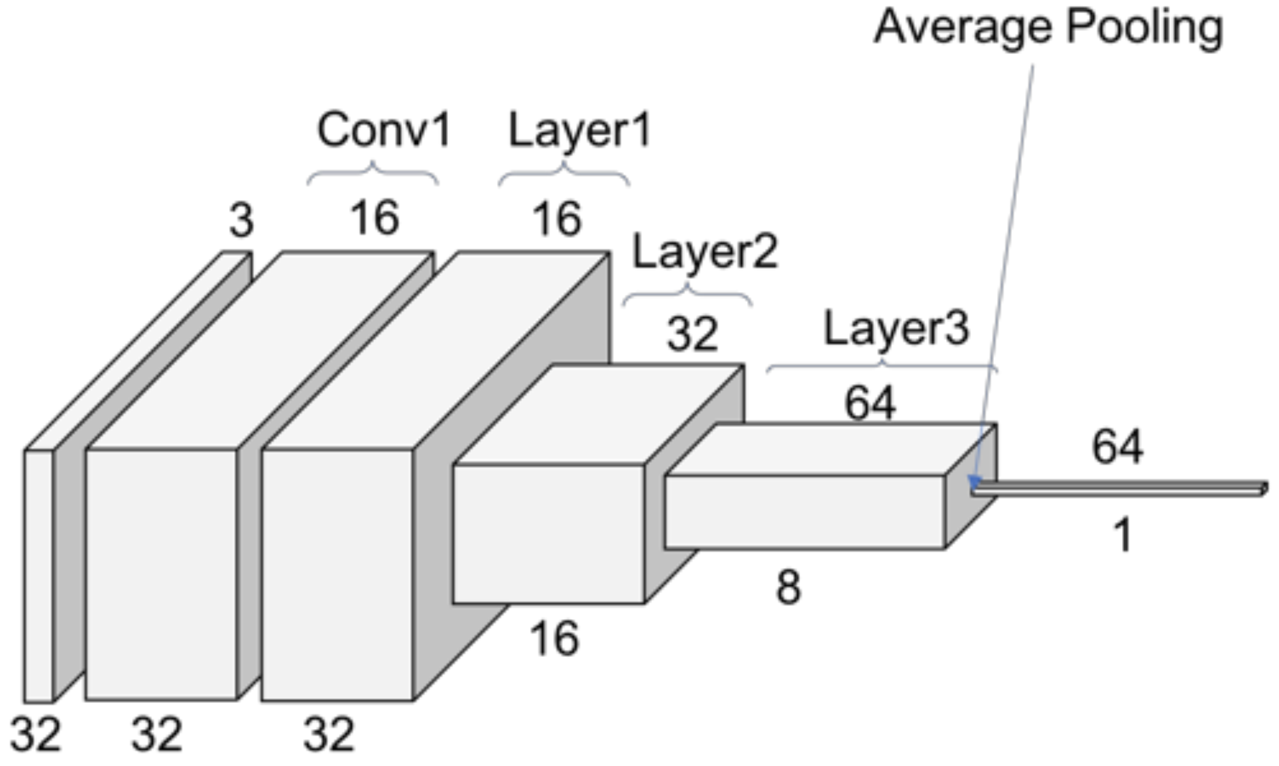
$p_{\theta^t}^{c}$ is the softmax probability of the class c using the network at time t,

K^{t-1} is the set of classes known at time t-1.

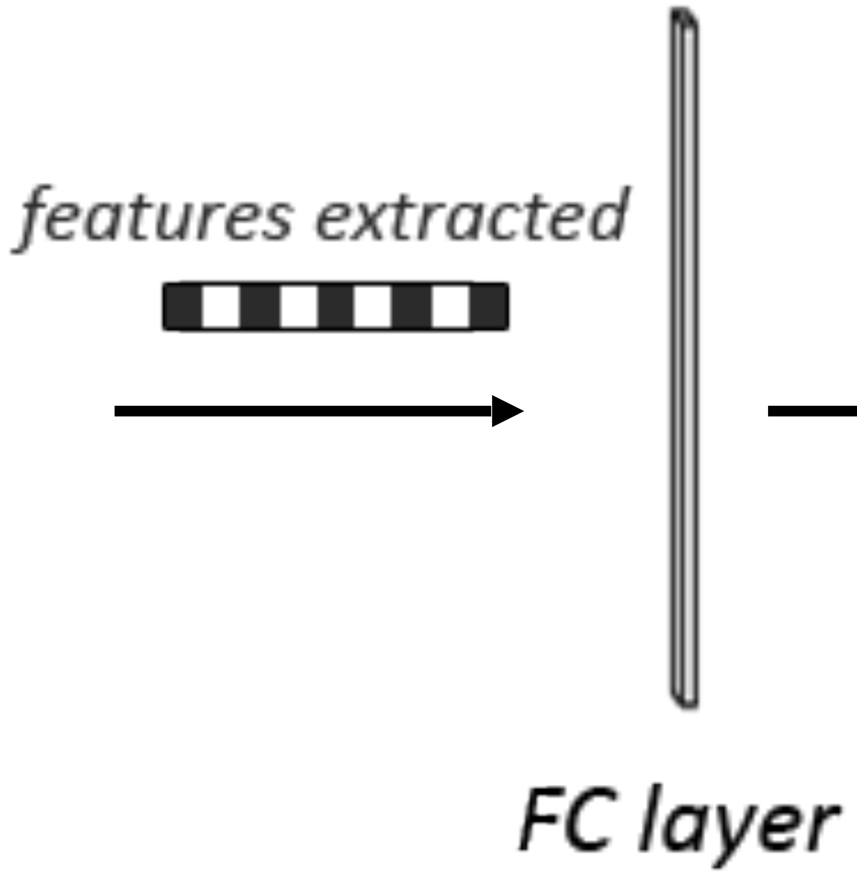
Intuition: Keep the same probability for previous classes to prevent forgetting.

Z. Li and D. Hoiem. Learning without forgetting. In IEEE T-PAMI, 40(12):2935–2947, 2017

iCaRL: Incremental Classifier and Representation Learning



$\gamma(x)$



Binary CE Loss

S. A. Rebuffi et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

iCaRL: Incremental Classifier and Representation Learning

Binary CE Loss

$$\ell_{bce}(x,y) = -\Big(\sum_{c \in Y^t} y_c \ \log p^c_{\theta^t} + (1 - y_c)\ \log(1 - p^c_{\theta^t})\Big) + $$

$$-\Big(\sum_{c \in K^{t-1}} p^c_{\theta^{t-1}} \log p^c_{\theta^t} + (1 - p^c_{\theta^{t-1}})\ \log(1 - p^c_{\theta^t})\Big)$$

$p^c_{\theta^t}$ is the binary probability (sigmoid) of the class c using the network at time t,

K^{t-1} is the set of classes known at time t-1, Y^t is the set of new classes.

S. A. Rebuffi et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

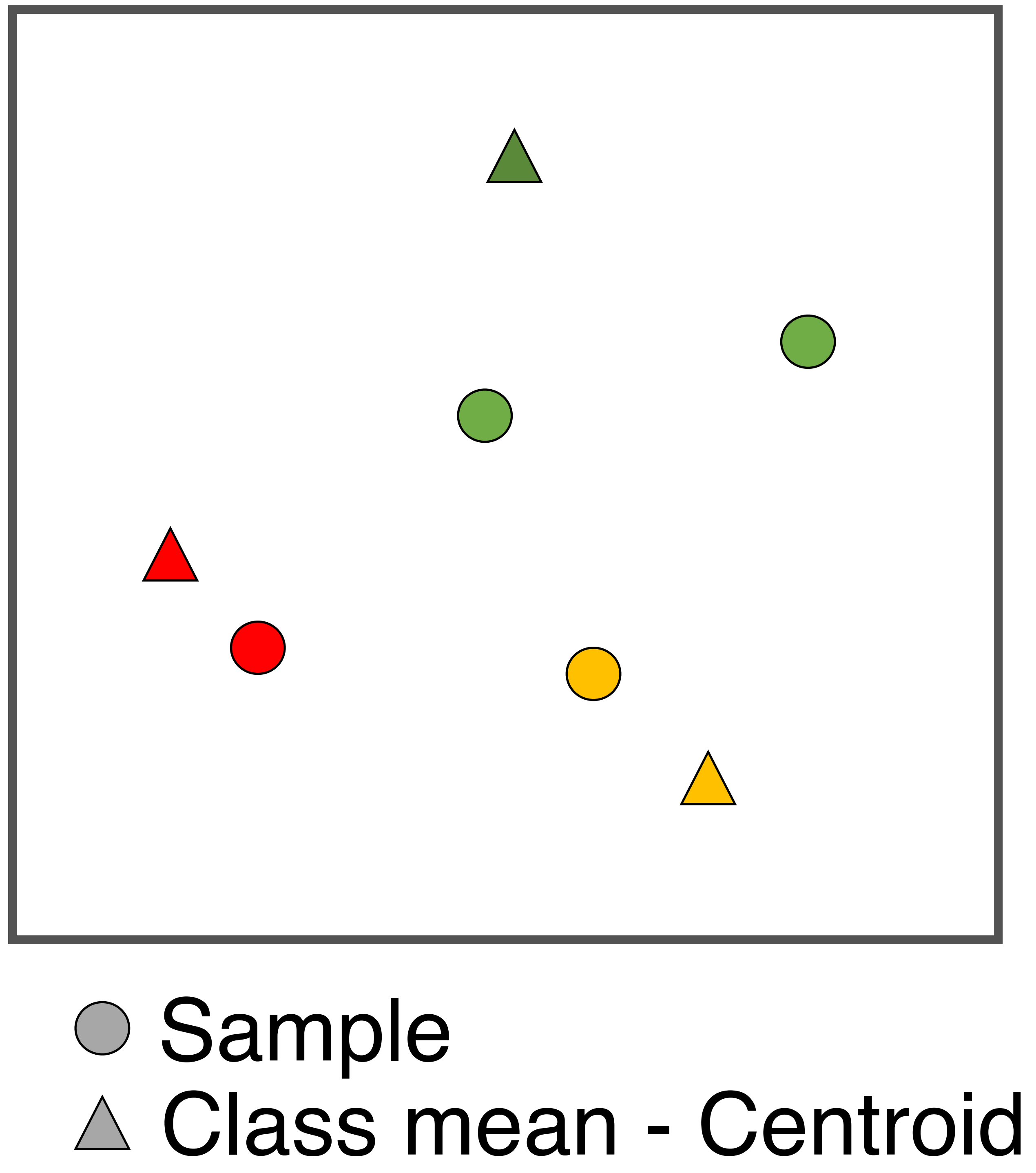
iCaRL: Incremental Classifier and Representation Learning

Nearest-Mean-of-Exemplars Classification (NME)

$$\bar{y} = \text{arg min}_c ||\gamma(x) - \mu_c||^2$$

$\gamma(x)$ are the features before the classifier layer
μ_c is the mean of features vectors of class c

Intuition: Pick the class for which the distance among the sample and the mean is minimal



S. A. Rebuffi et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

iCaRL: Incremental Classifier and Representation Learning

Exemplars

A memory containing samples of previous classes.

- Useful to rehearse knowledge of previous classes, reducing catastrophic forgetting.
- It can be used only a limited number of samples (usually 20 images per class).
- The use of exemplar is still debated in incremental learning community since they break the assumption that old data are no more available (but commonly used).
- Different strategies to select the best samples to store (often random selection works as well as more complex approaches)

S. A. Rebuffi et al. iCaRL: Incremental Classifier and Representation Learning. CVPR 2017

Other approaches

Many more strategies!

Many other approaches were defined for Incremental Class Learning:

- EWC [1], RW[2], SI [3] introduce a regularization term to prevent changes in “important" parameters.
- Progressive Neural Network [4] increases the size of the network adding novel neurons when extending the model knowledge.
- Bic [5] and LUCI [6] deal with the bias toward new classes proposing different ways to normalize the probabilities.
- DGR [7] and MerGAN [8] propose a *pseudo-rehearsal* approach: instead of memorizing samples, they employ a generative model to generate samples of previous classes.

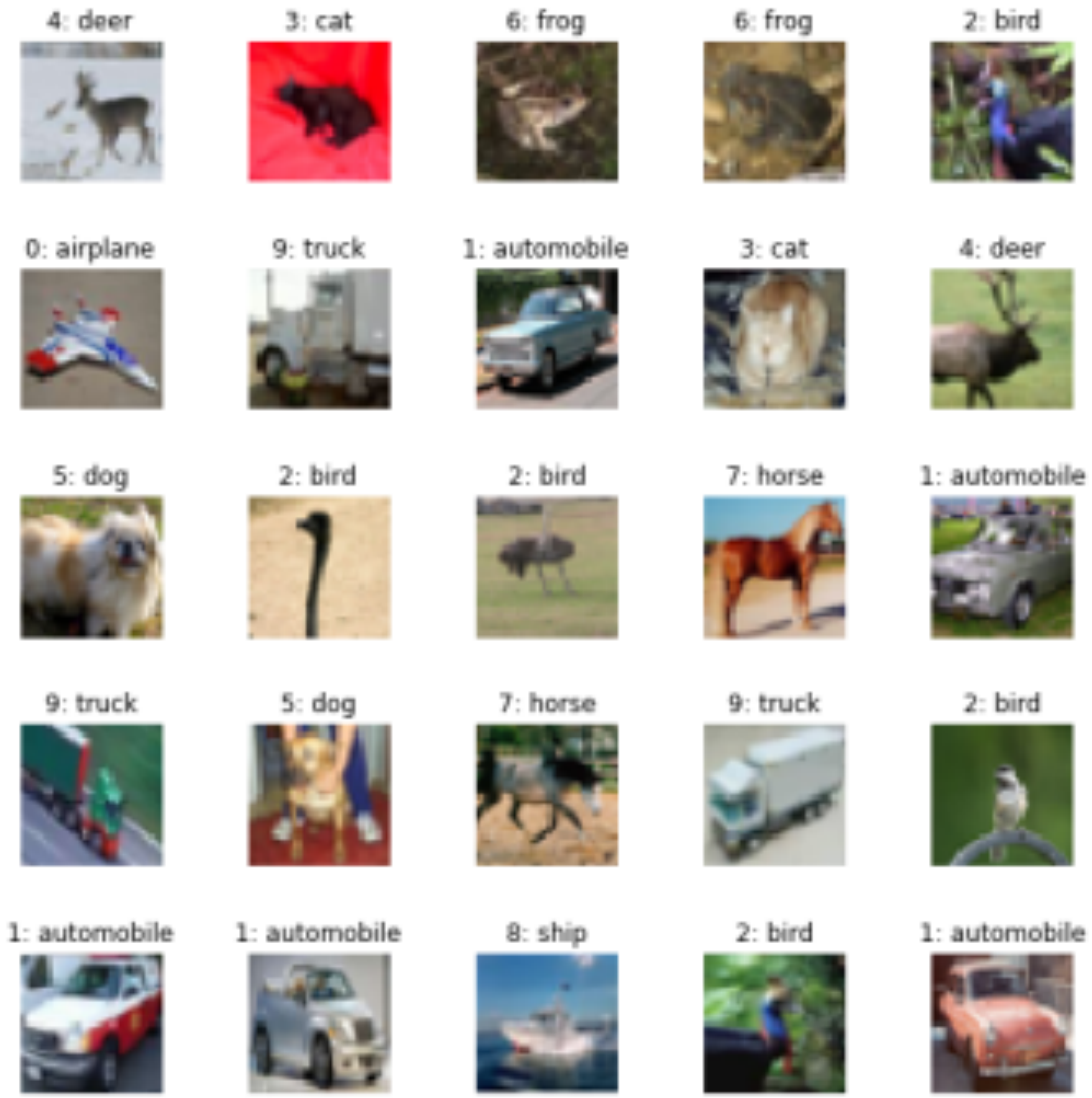
References are in the end of the slides.

Dataset - CIFAR 100

The dataset you will use for the project

CIFAR-100

Experiments



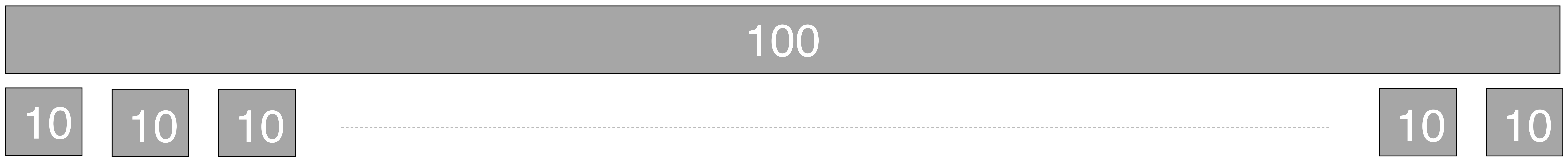
STATISTICS:

- *60.000 32x32 images*
- *100 classes*
- *50.000 training images*
- *10.000 testing images*

Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images", 2009.
url:https://www.cs.toronto.edu/~kriz/cifar.html

Learning procedure

Experiments



First, split the dataset in 10 groups of 10 classes

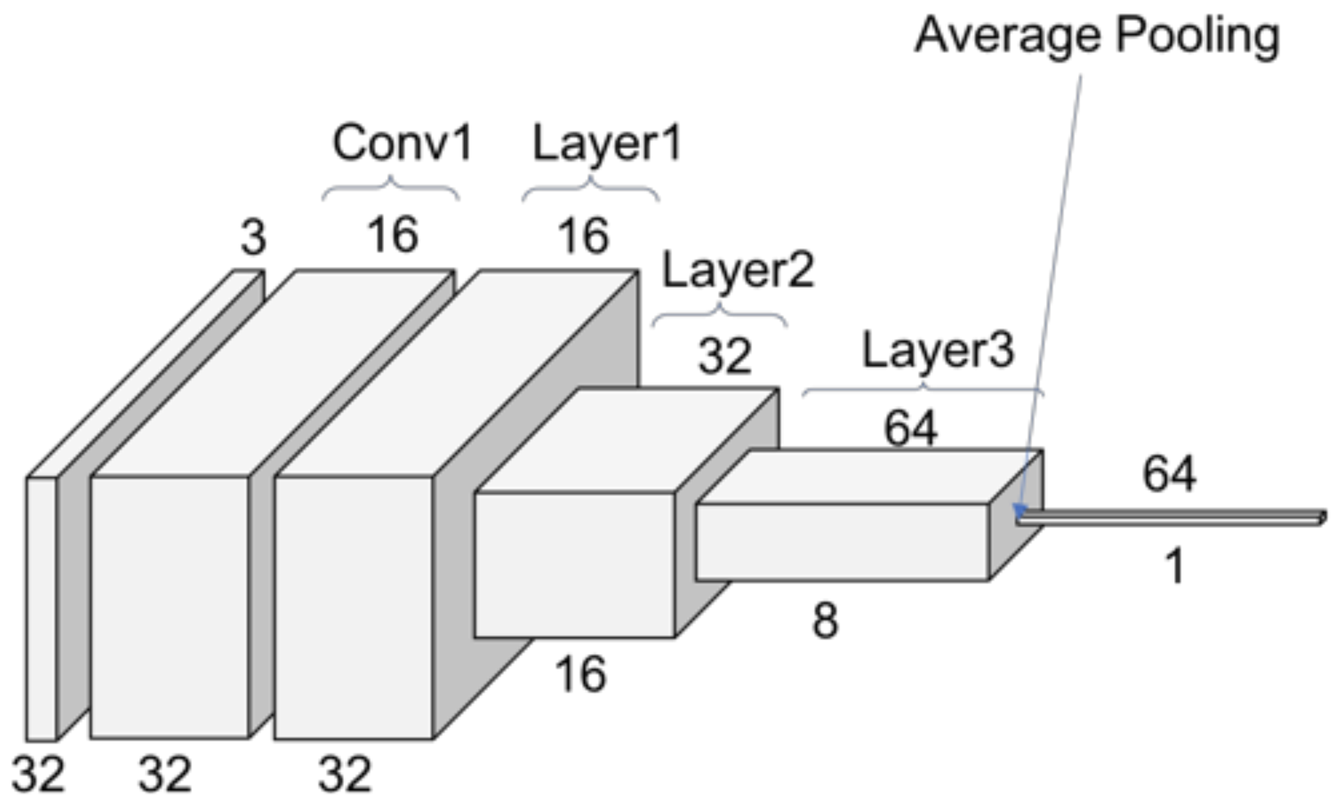
Learning procedure

Experiments

100

10 10 10 10 10

Start by learning the first one.

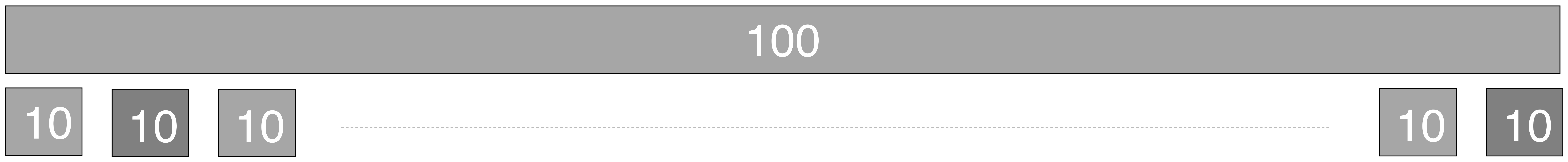


10

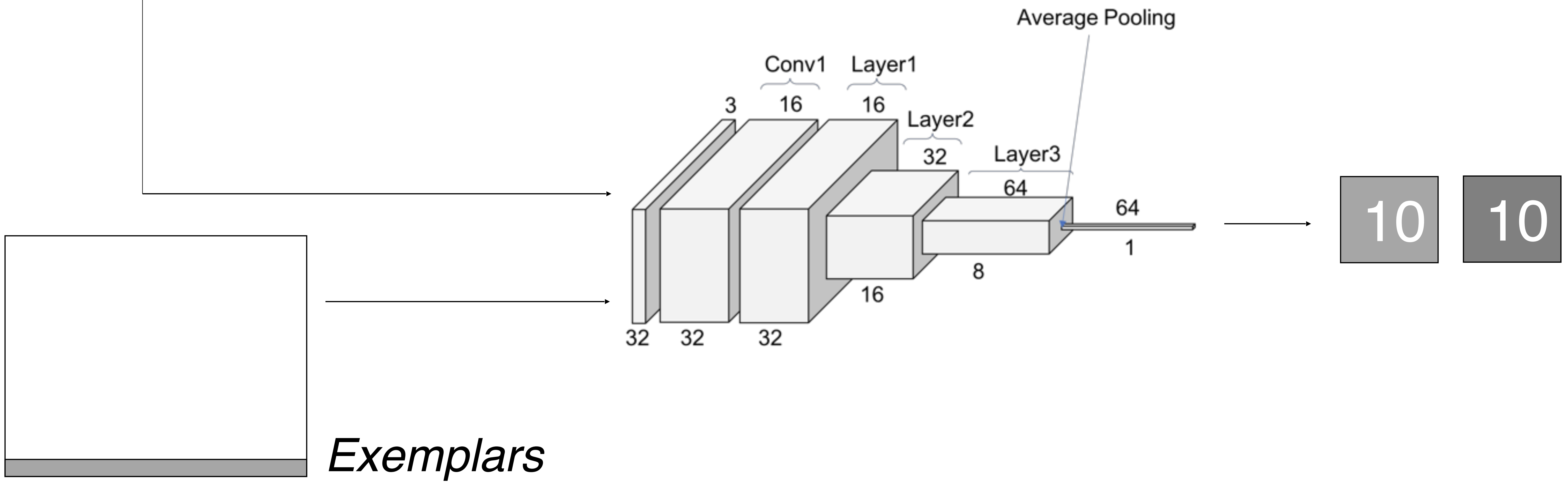
Exemplars

Learning procedure

Experiments

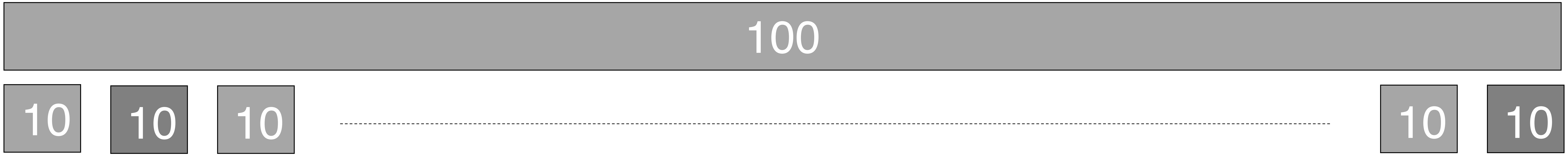


Then the second one….

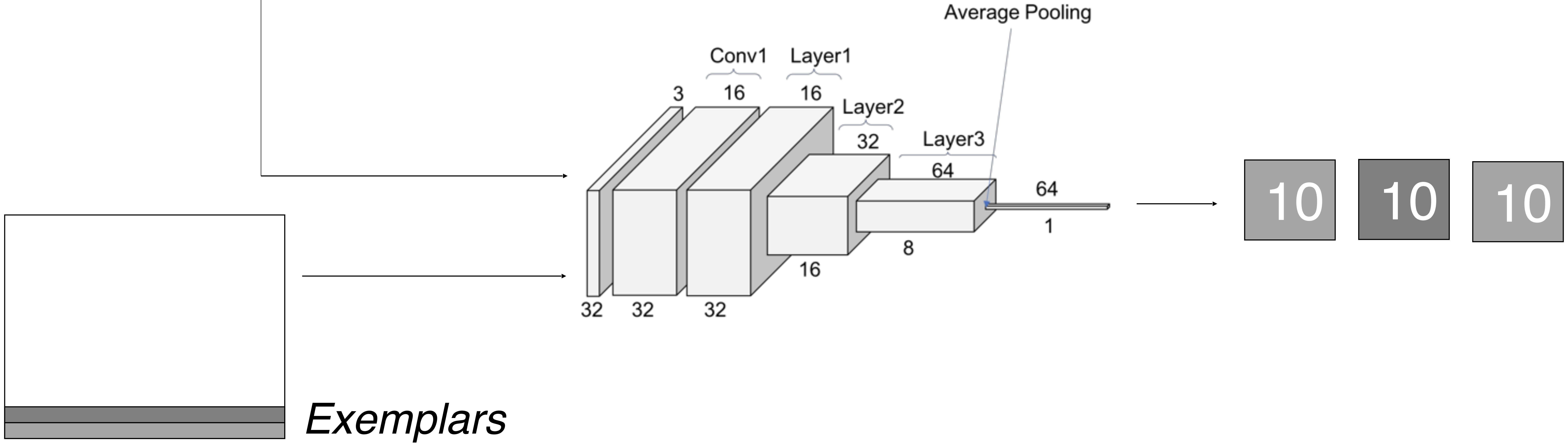


Learning procedure

Experiments

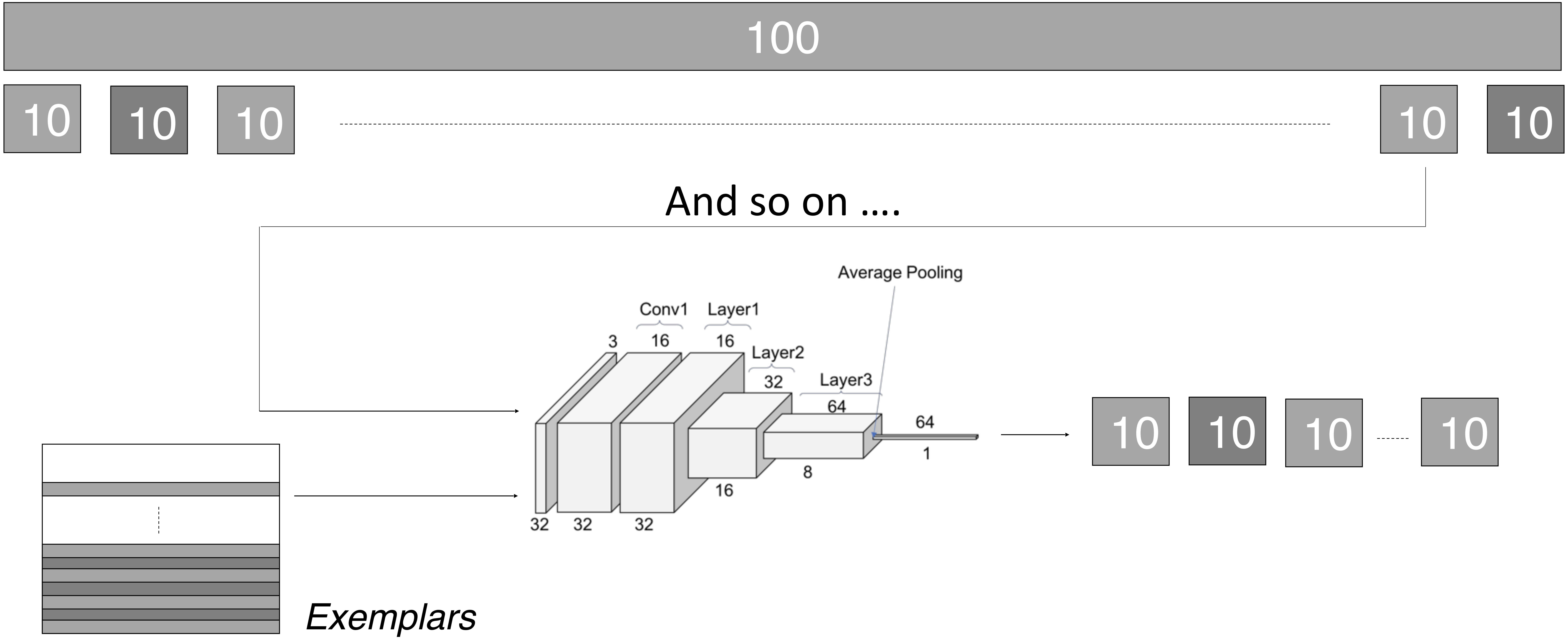


And the third one



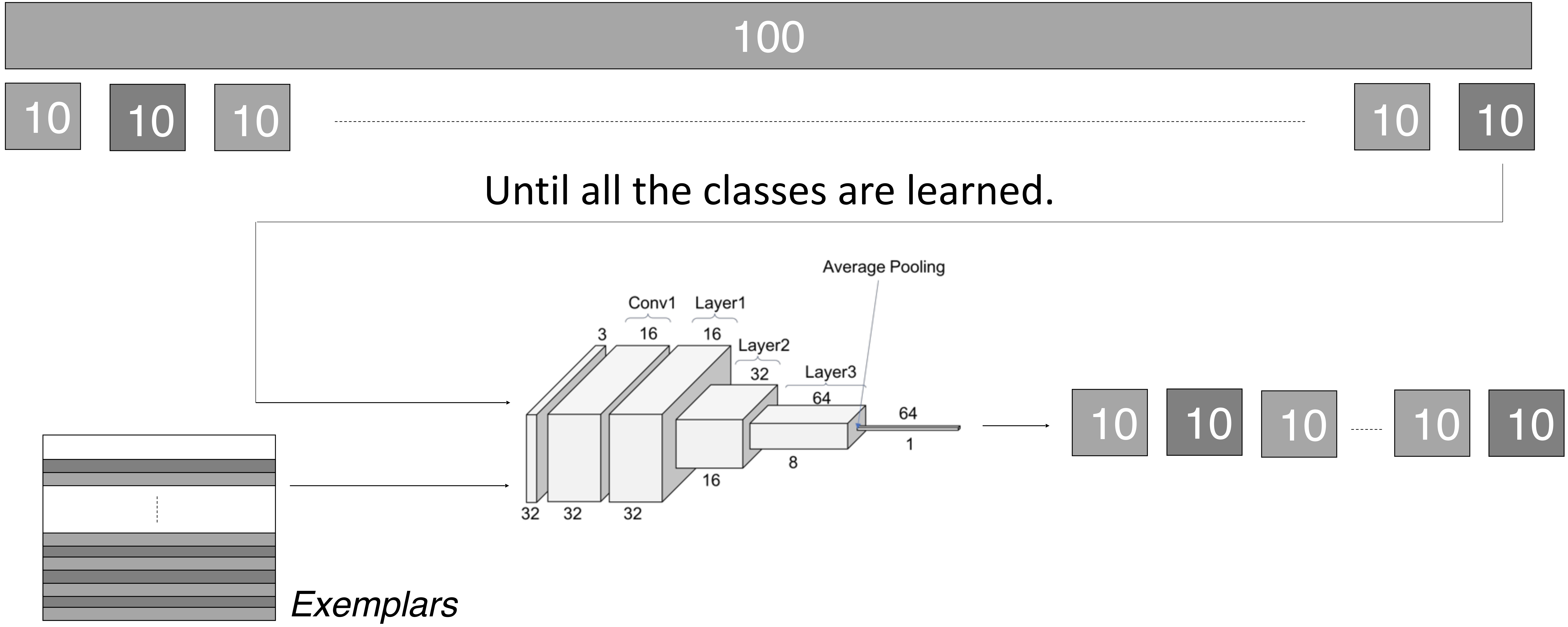
Learning procedure

Experiments



Learning procedure

Experiments



Testing protocol

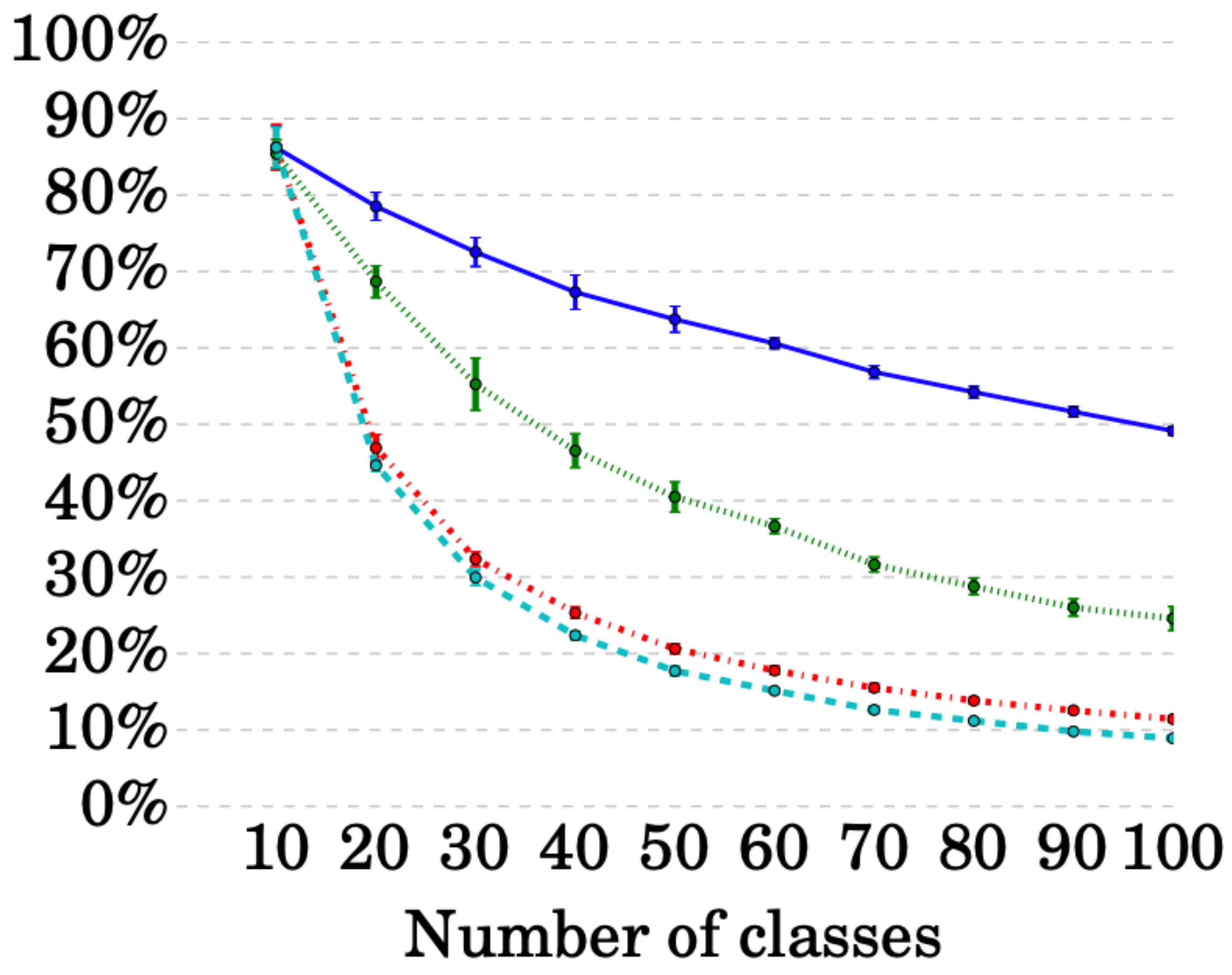
Experiments

100

10 10 10 ... 10 10

Test on all the classes seen so far.

Often reported as a plot.



Test 1 - 10 classes

Test 2 - 20 classes

Test 3 - 30 classes

Test 9 - 90 classes

Test 10 - 100 classes

Your project

What you need to do.

Where to Start

Baselines

1. Try out to learn new classes without using any presented strategy.
   - Fine-tune the network on the new classes
   - How does it affect the performance on old classes?

2. Implement Learning Without Forgetting and ICaRL.
   - Do they improve performance?
   - What are they weak points?

   (Use the setting and hyper-parameters presented in ICaRL)

Evaluate different choices

Ablation study

3. Are the losses for classification and distillation presented in LwF and ICaRL the best choices?

- Try out different combinations of classification and distillation losses.
- Which is the best one? Which is better in avoiding forgetting? Which is better in learning new classes?

4. Try different classifiers.

- Many different classifiers can be applied other than NME and a FC layer.
- Is there a more appropriate classifier than the one proposed?

It’s your turn!

Propose your own variation

5. After the previous steps, you should have identified the limitations of the current approaches. It’s your turn to propose an improvement!

- You can take inspiration from recent works but don’t limit yourselves only to incremental learning approaches.

- Try to think out of the box and feel free to change any aspects of the previous methods.

Have fun!

If something is not clear or you want additional information drop me an email!

fabio.cermelli@polito.it

Other approaches

Many more strategies!

EWC [1]: Overcoming catastrophic forgetting in neural networks, J. Kirkpatrick et al (Proceedings of the National Academy of Sciences 2017)

RW[2]: Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence, A. Chaudhry et al (European Conference Computer Vision 2018)

SI [3]: Continual Learning Through Synaptic Intelligence, F. Zenke et al (International Conference on Machine Learning 2017)

Progressive Neural Network [4]: Progressive Neural Networks, A. A. Rusu et al. (Not published - DeepMind)

Bic [5]: Learning a Unified Classifier Incrementally via Rebalancing, S. Hou et al (CVPR 19)

LUCI [6]: Large Scale Incremental Learning, Y. Wu et al (CVPR 19)

DGR [7]: Continual learning with Deep Generative Replay, H. Shin et al (NeurIPS 2017)

MerGAN [8]: Memory Replay GANs: learning to generate images from new categories without forgetting, C. Wu et al (NeurIPS 18)

Recent survey: Continual learning: A comparative study on how to defy forgetting in classification tasks, M. De Lange at al.