A Close Look at Deep Learning with Small Data

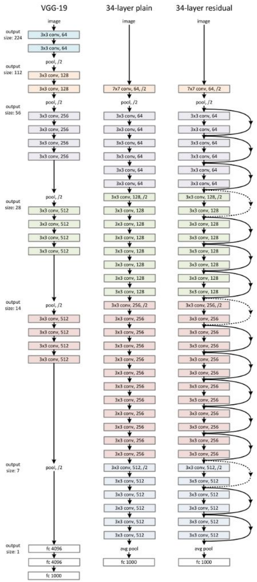
Lorenzo Brigato and Luca Iocchi

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

Deep Learning reached superior performance in many fields:

1. Lots of **data** (e.g. images, text)
2. High **capacity** neural networks (e.g ResNets)



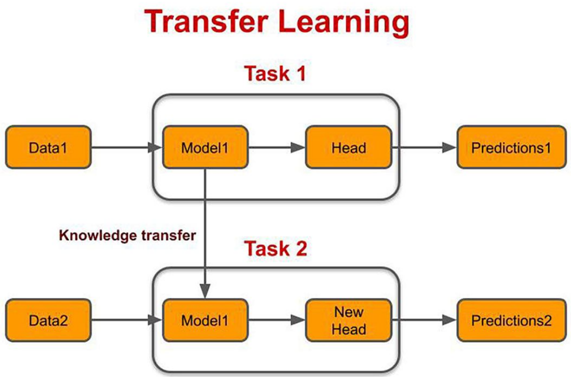
Problem:

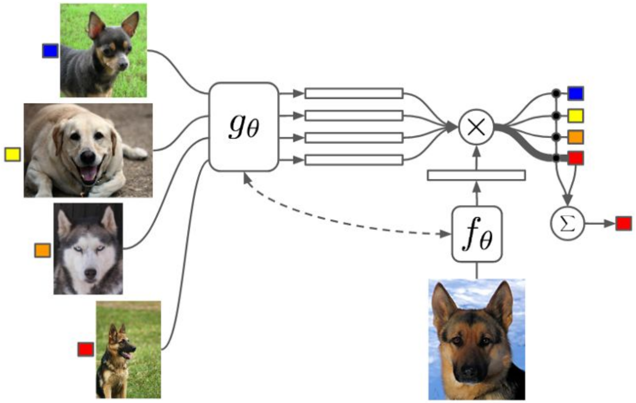
1. **Obtaining** data at large scales
   a. time-consuming
   b. difficult
2. **Labeling** data at large scales
   a. expensive

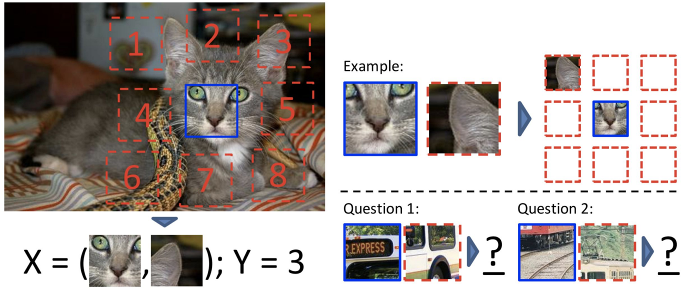
Introduction

Well-known approaches to decrease data needs (samples/labels):

1. **Transfer** learning
2. **Few-Shot** learning
3. **Self-Supervised** learning







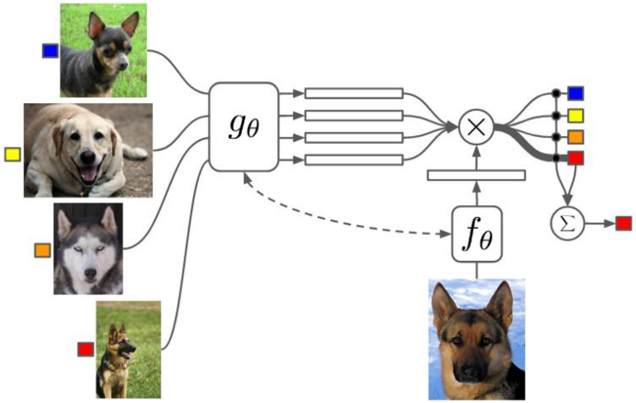
Introduction

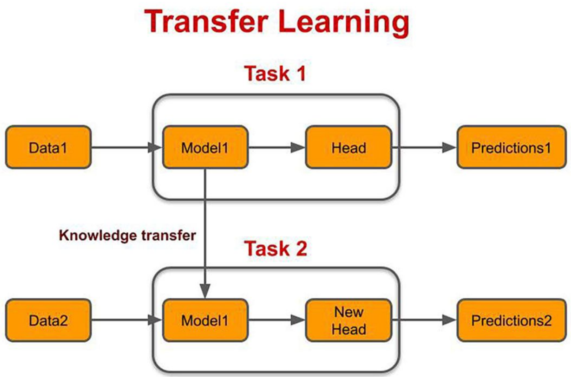
Well-known approaches to decrease data need (samples/labels):

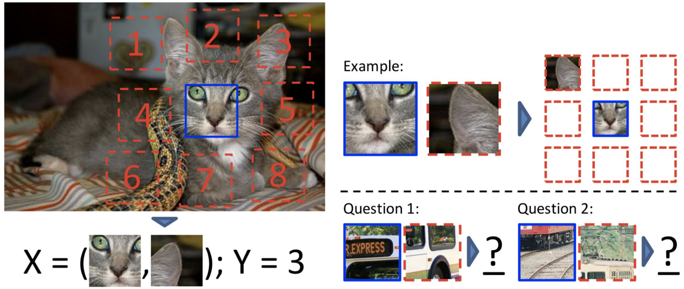
1. ~~Transfer learning~~
2. ~~Few-Shot learning~~
3. ~~Self-Supervised learning~~

Collect a large *source* labeled dataset

Collect a large *unlabeled* dataset







Problem Formulation

We are facing a supervised classification problem $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1) \ldots, (\mathbf{x}_s, \mathbf{y}_s)\}$

$\mathcal{D}$ is balanced and relatively small (constraining number of samples per class N)

No restriction on the number of classes K

Testing sets remain fixed at evaluation time

Objective $\mathbf{y} = f_\theta(\mathbf{x})$

In this work:

- $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$
- $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$

Related work

Vector data:

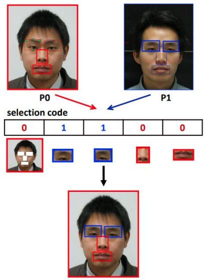
1. *Do we need hundreds of classifiers to solve real world classification problems?* [Fernandez et al. 2014]
2. *Modern neural networks generalize on small data sets* [Olson et al. 2018]

Random Forests and MLPs were the best models

Image generation:

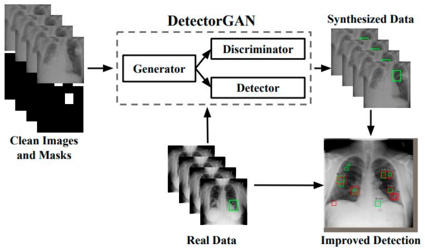
Frankenstein: Learning deep face representations using small data [Hu et al. 2017]

Facial recognition, very **domain specific**



Generative Modeling for Small-Data Object Detection [Liu et al. 2019]

CT images detection



Related work

Algorithmic approaches on image datasets:

1. *Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks* [Arora et al. 2020]

   Neural tangent kernels (NTK) on UCI repository and Few-shot Learning experiments

   Convolutional neural tangent kernels (**CNTK**) for **small CIFAR10** task **better** than **ResNet-34**

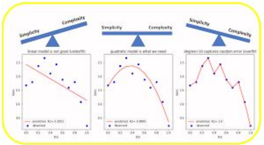
2. *Deep Learning on Small Datasets without Pre-Training using Cosine Loss* [Barz et al. 2020]

   Propose the use of **cosine** loss instead of **cross-entropy** loss

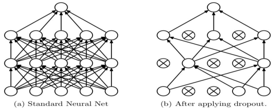
   Improved results mainly on **fine-grained datasets** with 20 - 80 samples per class and 66 - 555 classes

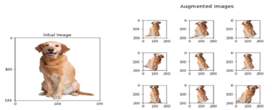
Empirical study

1. **Influence** of **model complexity** on performance
    a. CNN with 4 conv layers, 16/32/64 base filters (CNN-lc/mc/hc)
    b. ResNet-20 with 16 base filters



2. **Influence** of **regularization** techniques on performance
    a. Dropout with varying drop-rates (0.0/0.4/0.7)
    b. Enable/disable standard data augmentation
        i. Cropping + flipping on CIFAR-10
        ii. Cropping + flipping + color distortion on SVHN
        iii. Cropping on FMNIST



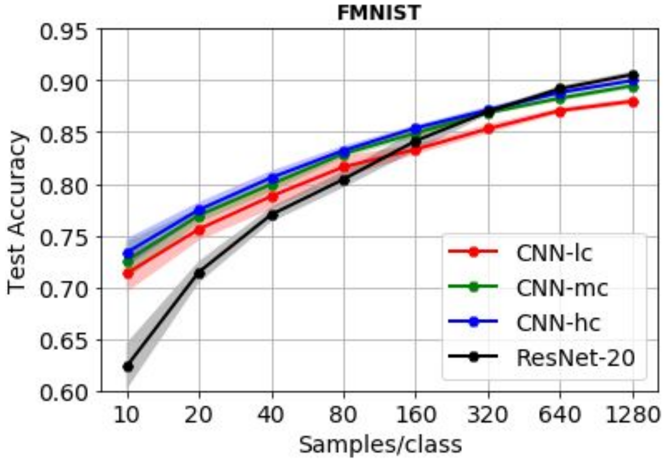


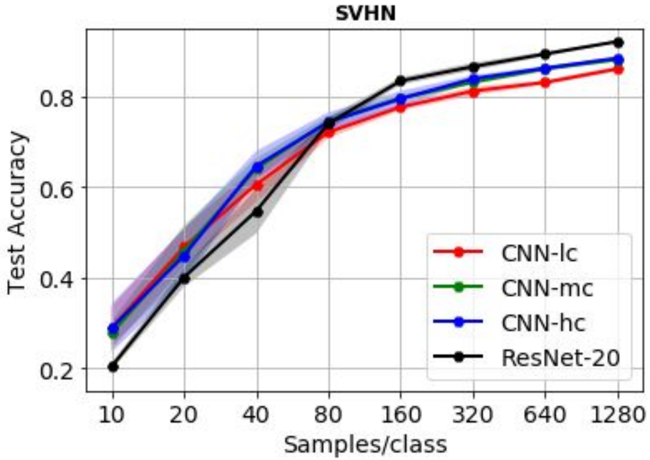
3. **Comparison** of baseline models with **state-of-the art** approaches:
    a. CNTK [Arora et al. 2020]
    b. Cosine loss [Barz et al. 2020]

Standard optimization set-up

1. *Adam* - default parameters for CNNs
2. *SGD* + *Nesterov* + *weight decay* = 1e-4 + piecewise learning rate schedule for ResNet
3. Epochs changed according to the size of model and training set
4. Batch size = 32
5. Cross-entropy loss

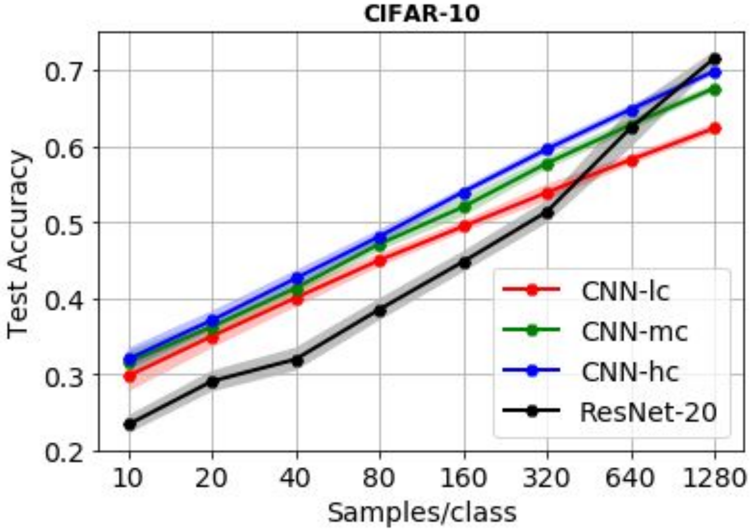
Results



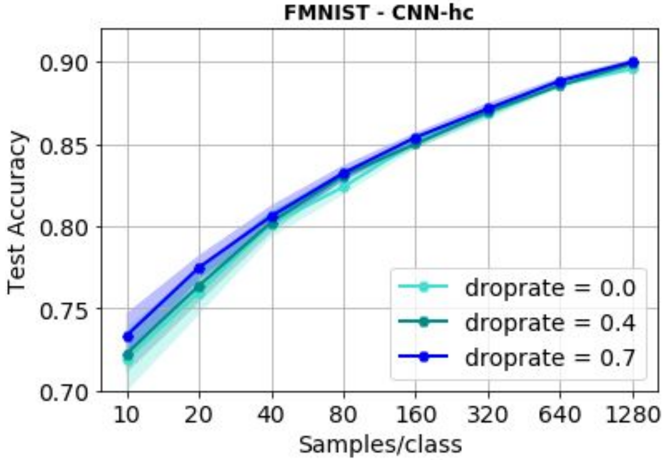


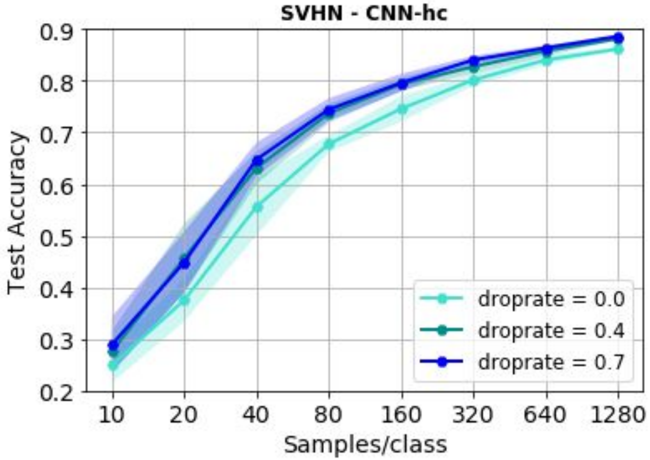
Model complexity is a critical factor:

- Basic CNNs -- small datasets
- ResNet -- larger datasets



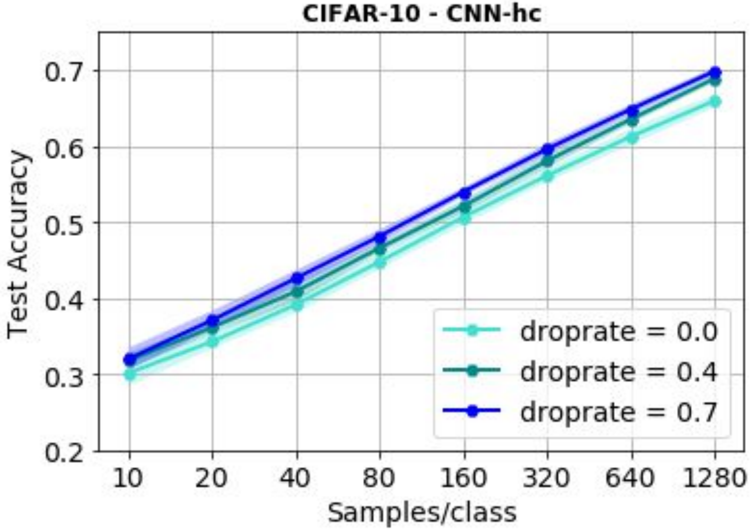
Results



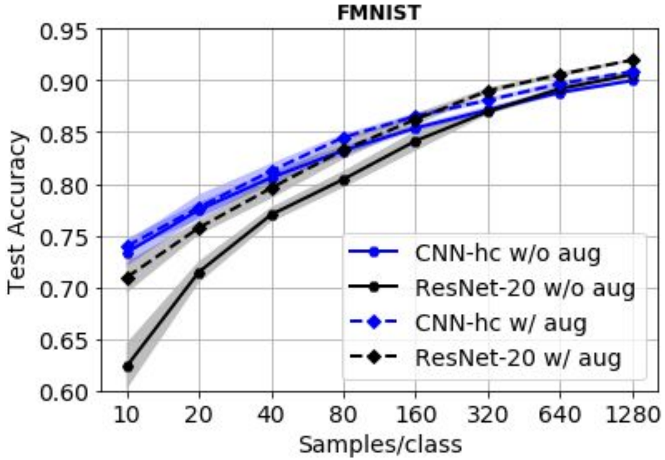


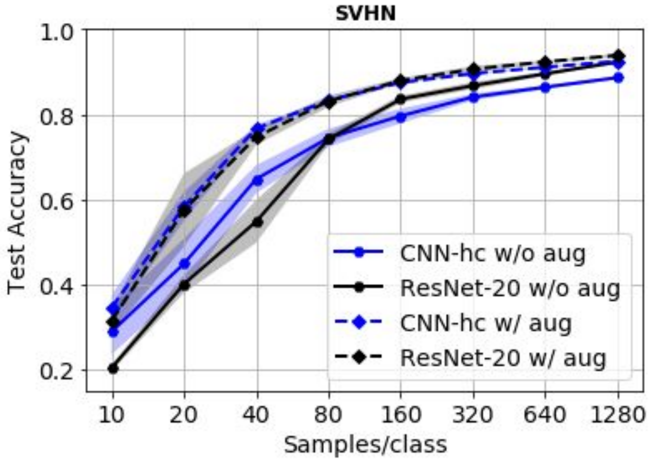
Dropout remains a good regularizer:

- Gains up to 10%



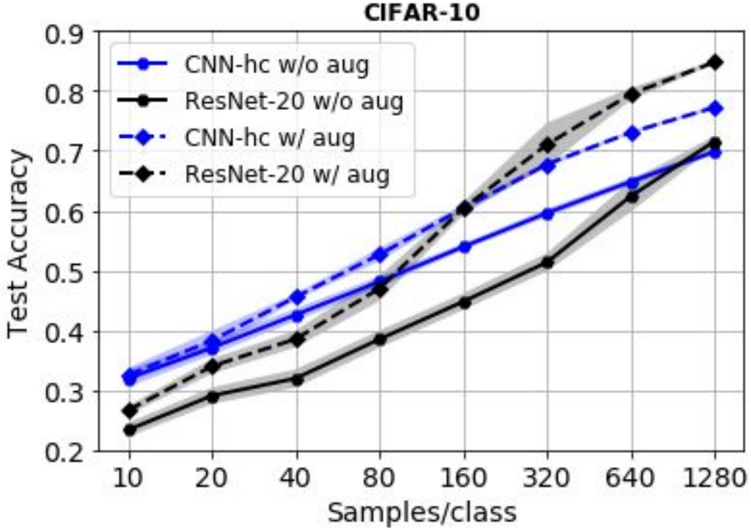
Results



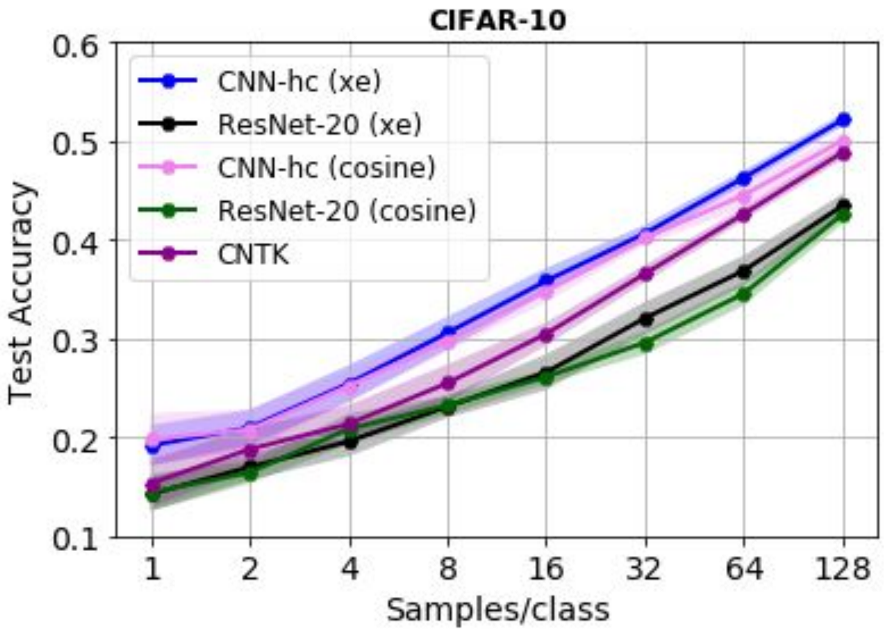


Data augmentation can induce large gains:

- Up to 20% (ResNet-20)
- Up to 10% (Cnn-hc)



Results



Comparison with state of the art:

- **ResNet-20 struggles** with extremely **limited data**
- **Cosine** and **cross-entropy** losses are **comparable**
- Basic **CNN-hc outperforms** the **CNTK** [Arora et al. 2020] by up to 5%

Considering evaluation protocol of [Arora et al. 2020] (no data augmentation)

Conclusions

Model complexity is a critical factor for small data domains when using standard training set-ups:

1. New proposed models should be compared to simple nets as well

Regularization:

1. **Dropout** is a good **regularizer** also with small data
2. Even standard data augmentation can induce large gains:
   a. Most promising direction

Baseline models are **better** than or **comparable** to state-of-the-art approaches in the tested set-up:

1. **Cross-entropy** loss is **comparable** to the **cosine** loss
2. A **shallow CNN** with 64 base filters **outperforms CNTK**