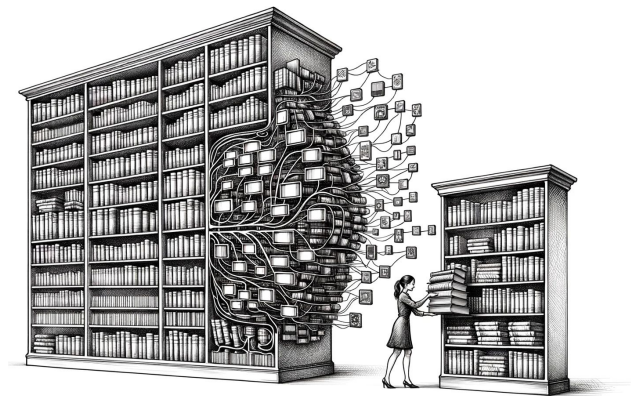




University of St.Gallen

Institute of Computer Science

Make Deep Networks Shallow Again



From insight
to impact.

Bernhard Bermeitinger, Tomas Hrycej, Siegfried Handschuh
KDIR2023, Nov 13th 2023, Rome

Outline

1. Motivation

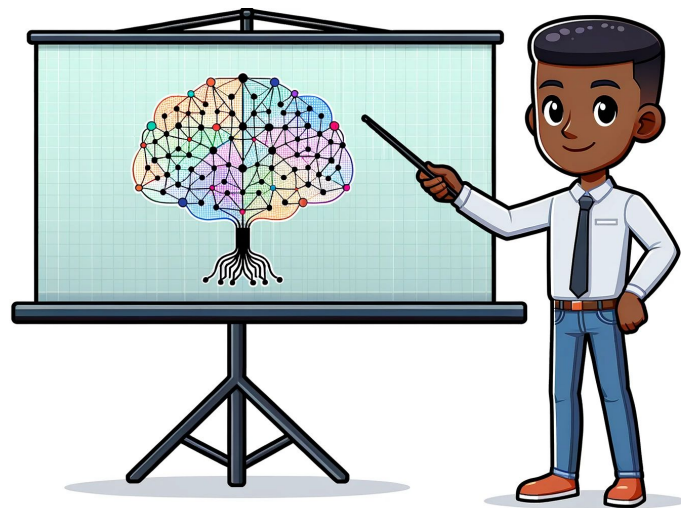
- a. Sequential CNN
- b. Residual Connections?
- c. Parallel CNN

2. Experimental Setup

3. Results

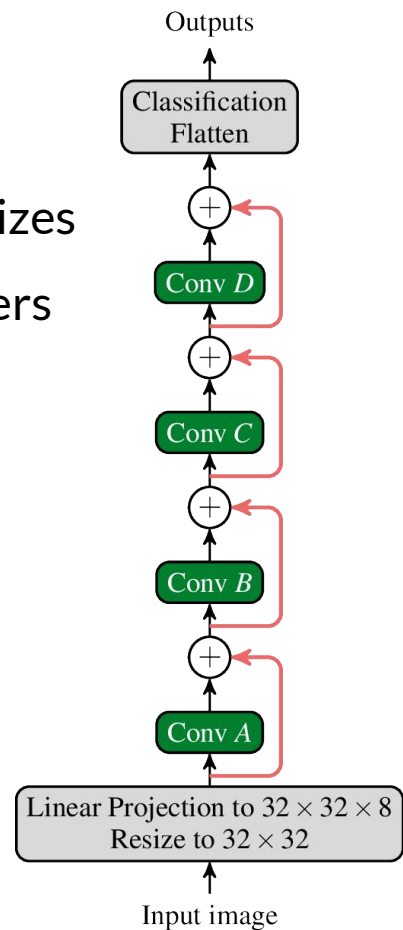
- a. 1 filter, 1-32 layers
- b. 1-32 filters, 16 layers

4. Summary



Sequential model

- Resize images to 32×32 for sane possibilities of filter sizes
- Project input to same dimensionality as number of filters
- No pooling to keep output dimensions equal
- Small classification layer for use case
- Skip connections **for each** convolutional layer

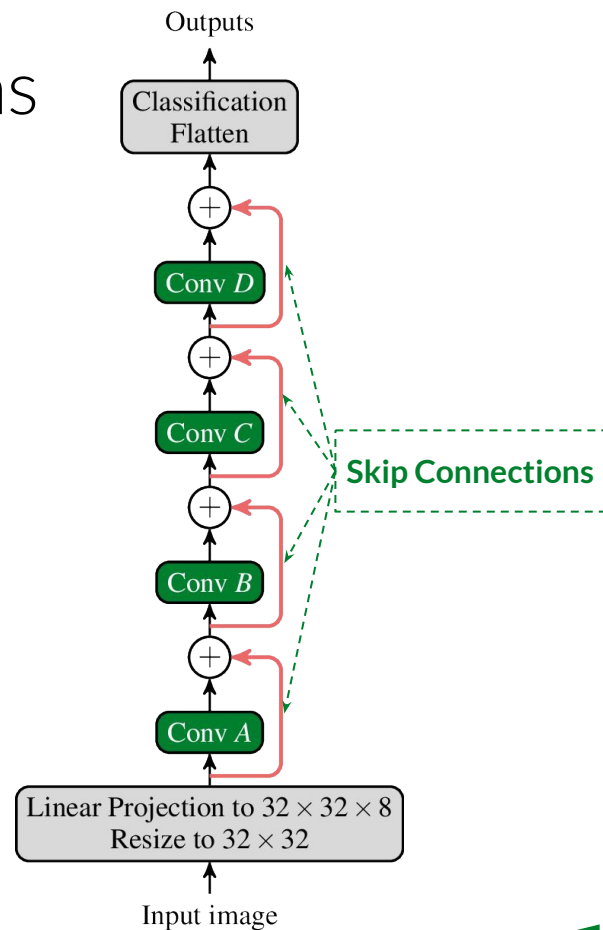


Motivation against residual connections

- CNNs increase in depth
 - Leads to **vanishing gradient**
- Solution: **Residual Connections**
 - Layer is skipped, input is **added** to layer output
 - Assumption: layer only needs to transform data marginally

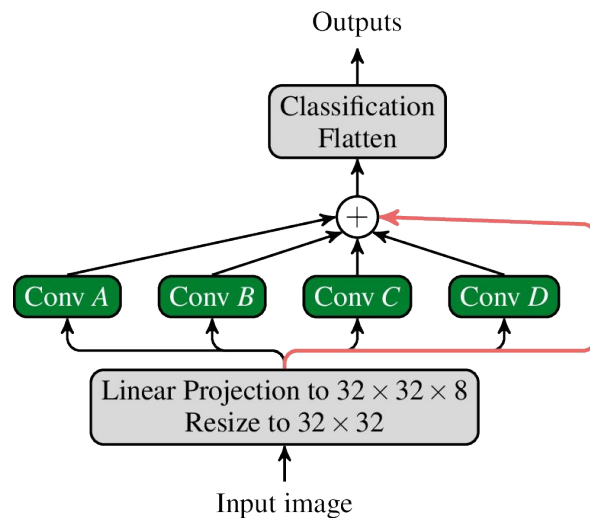
- Question:

*“If the later layers only result in small changes,
why even have them?”*



Parallel model

- Change topology to a parallel structure
- Approximation of sequential model
Linearization of gradients
- Each layer receives same input
Outputs are summed
- Acts like a **shallow** network with one layer
Less numerical instability / vanishing gradient



The (over-)determination ratio Q

K : Number of training examples

M : Number of classes

P : Number of trainable parameters

Overdetermination Ratio $Q = \frac{KM}{P}$

With $Q < 1$: More *parameters* than *constraints* $\Rightarrow$ underdetermined system

With $Q > 1$: More *constraints* than *parameters* $\Rightarrow$ **overdetermined system**

Desired state! No arbitrary parameters (no overfitting)

For CIFAR10: $50,000 * 10 = 500,000$ constraints

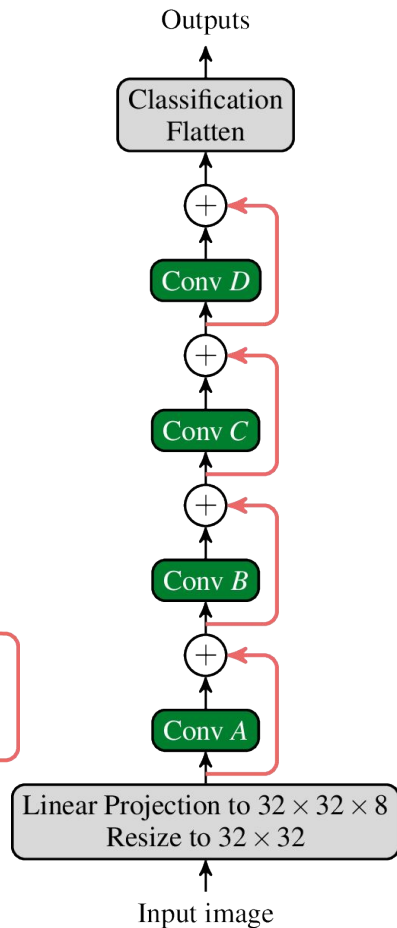
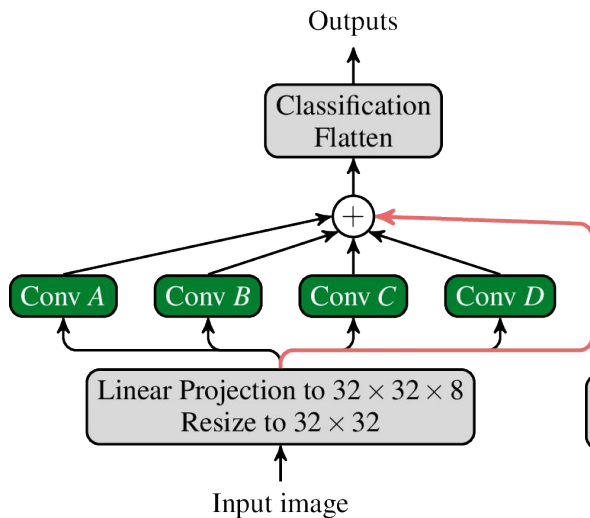
Perfect fit: 500,000 parameters

Definitely overfitting: $> 500,000$ parameters

Desired: 50,000 – 250,000 parameters

Experiment setup

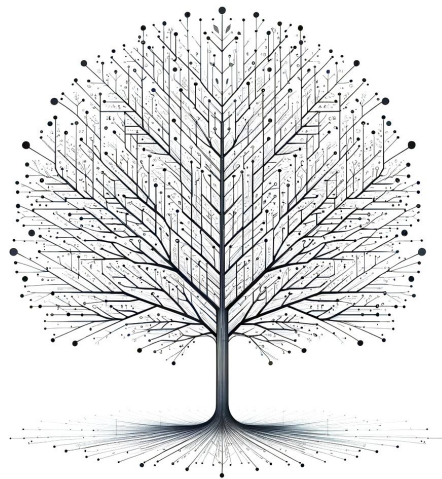
- Randomness important to pin to same seed
 - E.g. both Conv A layers initialized equally
- Datasets: MNIST and CIFAR10
 - Resized to 32x32
 - No augmentation
 - Projected to filter size
 - Metric: Cross-Entropy Loss
(not accuracy, it's not what is trained)
- Training setup
 - Trained for exactly 100 epochs
 - Guaranteed same batches of size 512
 - RMSprop as optimizer



Experiment variables

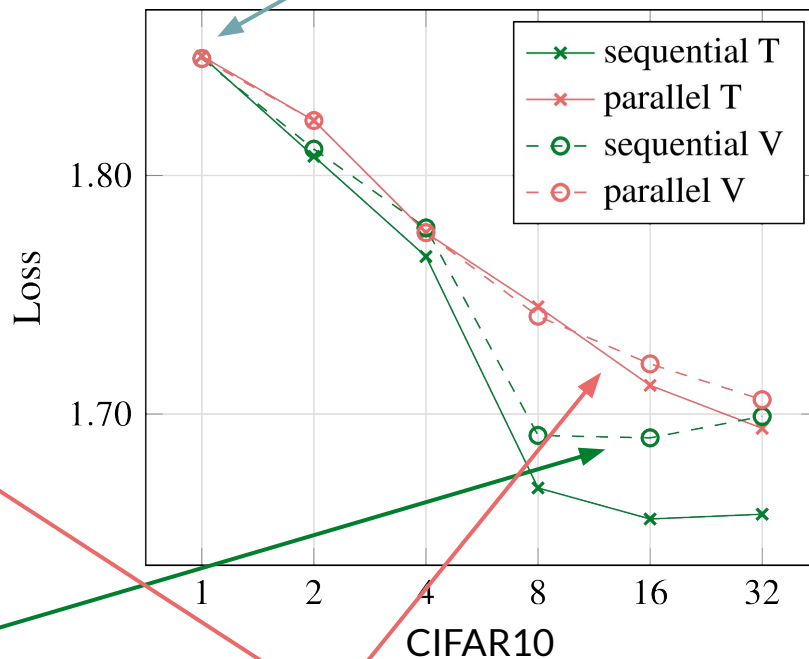
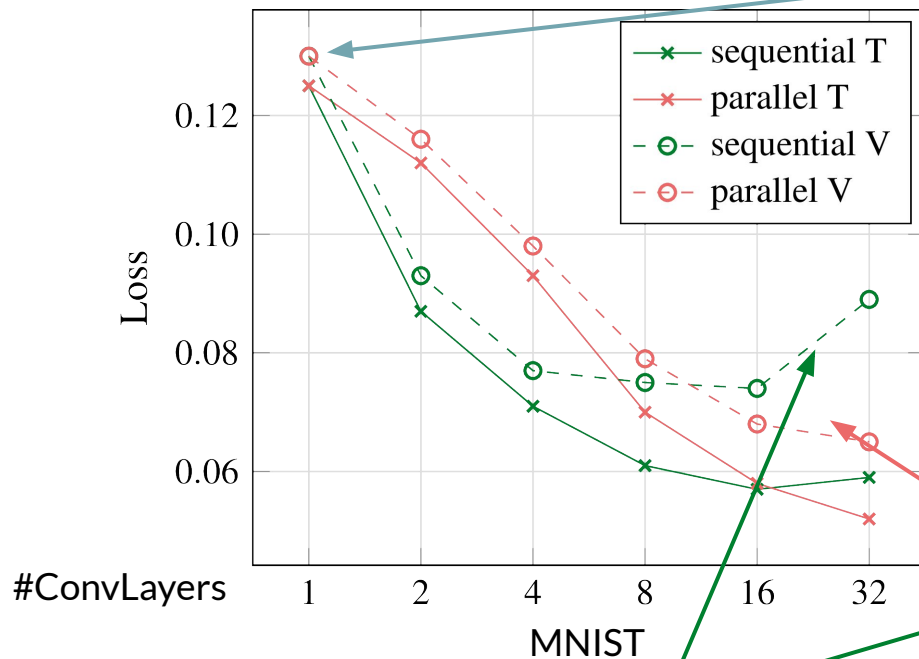
- 2 models: parallel, sequential
- 2 datasets: MNIST, CIFAR10
- 6 number of layers: 1, 2, 4, 8, 16, 32
(With 1 layer, the two topologies are equal)
- 6 number of filters: 1, 2, 4, 8, 16, 32
- 6 sizes for filters: 1, 2, 4, 6, 8, 16
- 2 activation functions: sigmoid, ReLU
- 4 learning rates: 10^{-2} , 10^{-3} , 10^{-4} , 10^{-5}

- 6,912 experiments
- ~80 GPU days



1 filter vs 1-32 layers

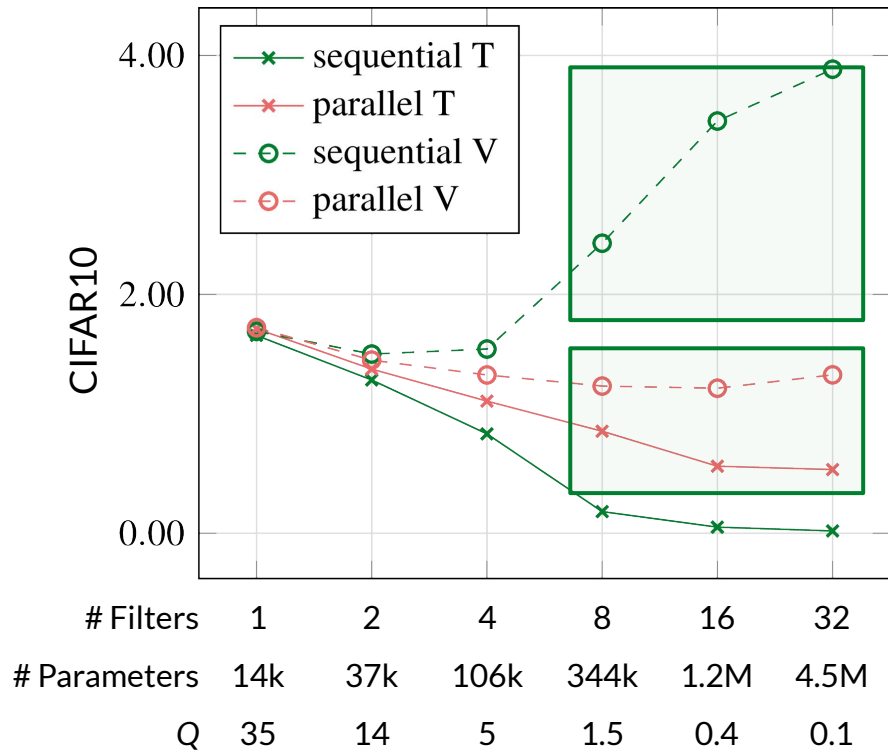
1 layer $\Rightarrow$ same topology $\Rightarrow$ exactly same results



Overfitting for sequential models

No overfitting for parallel models

1-32 filters, 16 layers



Massive overfitting

Slight overfitting

Definitely overfitting: > 500,000 parameters
Desired: 50,000 – 250,000 parameters

Summary

- Deep Residual CNNs can be condensed to shallower architectures with similar representational capacities
- Deep Networks tend to overfit due to capturing abrupt nonlinearities
- Shallow Networks show better resistance to training noise

