TRAINING VERY DEEP NETWORKS

GROUP 02

Date of Presentation : 29-09-2021

E/16/022 Chamath Amarasinghe (e16022@eng.pdn.ac.lk)

E/16/025 Diwanga Amasith (e16025@eng.pdn.ac.lk)

E/16/222 Wishwa Madushanka (e16222@eng.pdn.ac.lk)

E/16/232 Shamra Marzook (e16232@eng.pdn.ac.lk)

Background Details

Authors :

Professor Jürgen Schmidhuber - University of Lugano, Switzerland

Klaus Greff - University of Lugano, Switzerland

Rupesh Kumar Srivastava - University of Lugano, Switzerland

Published in :

International Conference on Machine Learning (ICML) 2015 (+600 citations)

Conference location and date :

Lille, France, 6 - 11 July 2015

Abstract

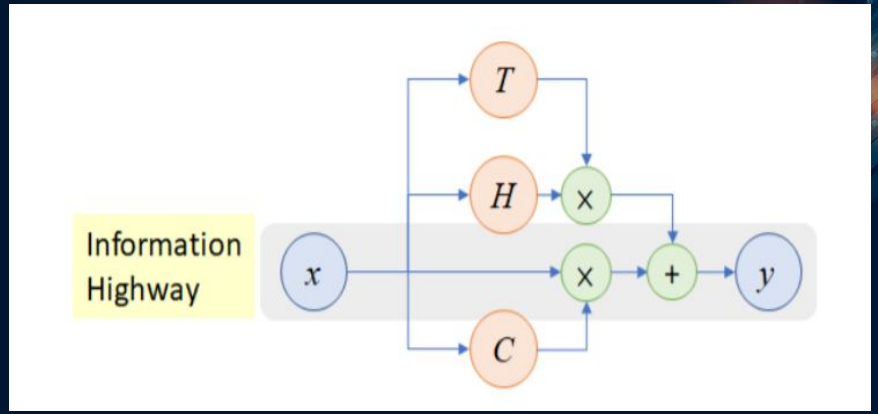
- Depth of neural networks - crucial
- Empirical breakthroughs in supervised machine learning - most important role
- Training - more difficult as depth increases, & remains an open problem
- Inspired by LSTM recurrent networks - highway networks are created
- Highway networks can be trained simple gradient descent.
- This enables the study of extremely deep and efficient architectures.

Highway Network

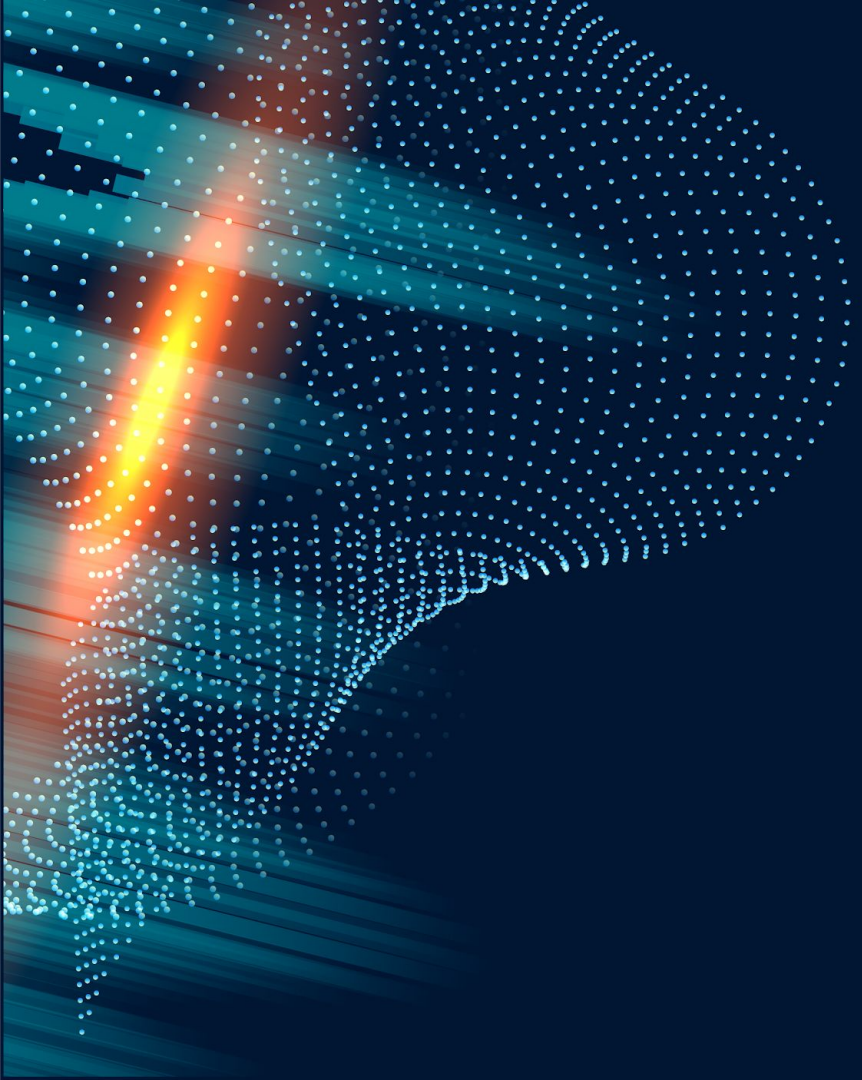
In highway network, two non-linear transforms T and C are introduced

y = H(x, WH). T(x, WT) + x. C(x, WC)

Where T is the Transform Gate and C is the Carry Gate



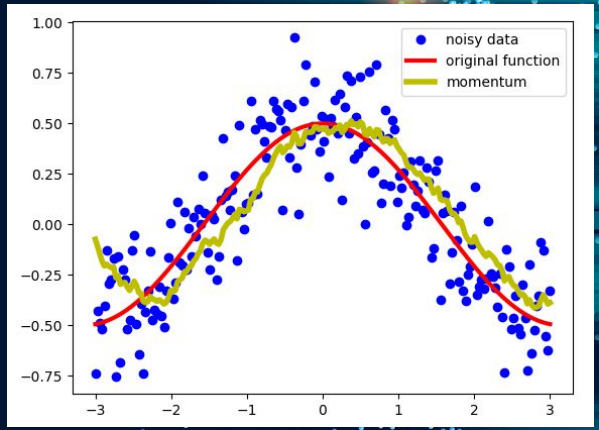
https://miro.medium.com/max/700/1*qHf_AHv8yJJsKQok4KS4Jw.png



EXPERIMENTS

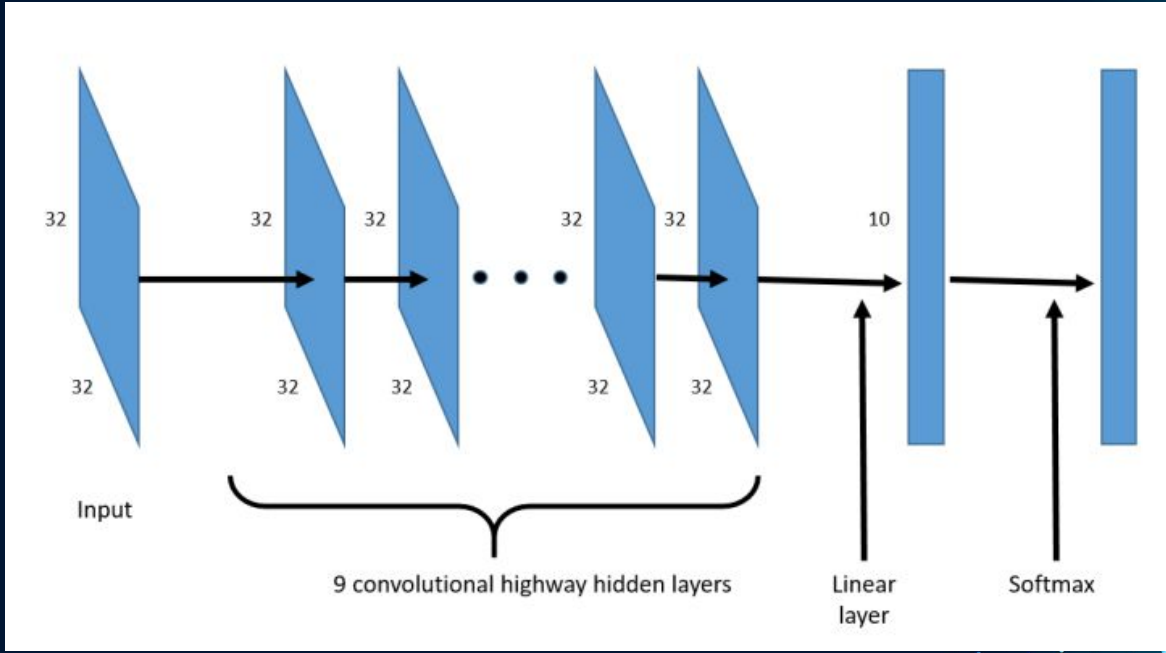
- All networks were trained using Stochastic Gradient Descent with momentum
- An exponentially decaying learning rate was used in optimization
- Highway networks utilize - ReLU activation function to compute H.
- Better estimate of the variability of classification

  Best (mean+- std.dev.)
- Frameworks used - Caffe and Brainstorm

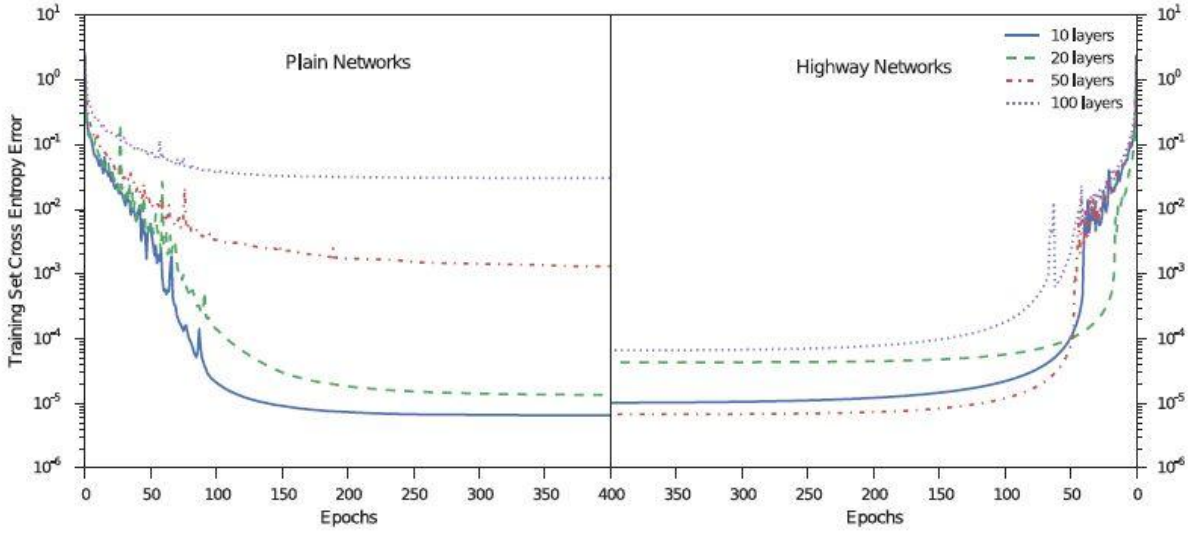


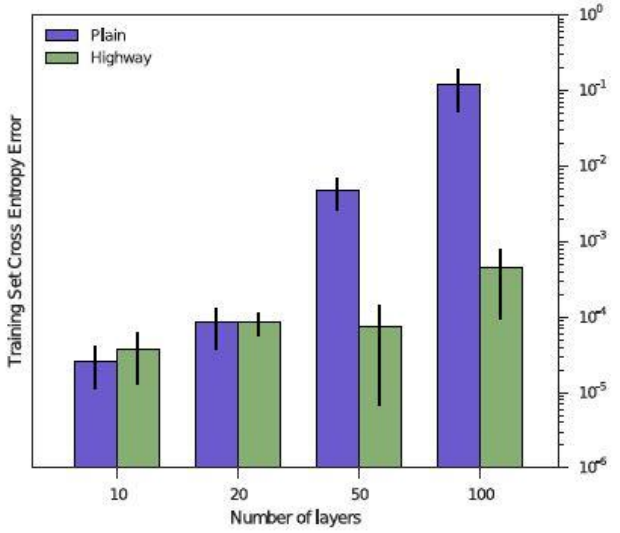
https://towardsdatascience.com/stochastic-gradient-descent-with-momentum-a84097641a5d

Optimization



Comparison of optimization of plain networks and highway networks of various depths.





E/16/232 Shamra

Experiments on MNIST Digit Classification

- MNIST database is collection of handwritten digits.
  - training set of 60,000 examples
  - test set of 10,000 examples.
- Trained 10-layer convolutional highway networks on MNIST
  - 9 convolutional layers, softmax output
  - filter maps (width) was set to 16 and 32 for all the layers.

Network	Highway Networks		Maxout [20]	DSN [24]
	10-layer (width 16)	10-layer (width 32)		
No. of parameters	39 K	151 K	420 K	350 K
Test Accuracy (in %)	99.43 (99.4±0.03)	99.55 (99.54±0.02)	99.55	99.61

The CIFAR-10 and CIFAR-100

- 80 million tiny images dataset
- CIFAR-10
  - Consists of 60000 32x32 colour images in 10 classes, with 6000 images per class
  - 50000 training images and 10000 test images.
- CIFAR-100
  - 100 classes containing 600 images each
  - 500 training images and 100 testing images per class
  - 100 classes are grouped into 20 superclasses

Comparison to Fitnets

- Maxout networks can cope much better with increased depth
  - Training on CIFAR-10 through plain back propagation only possible for maxout networks with a depth up to 5 layers
  - Having 5 layers limits number of parameters to 250K and the number of multiplications to 30M.
- Training of deeper networks was only possible through the use of a two-stage training procedure and addition of soft targets

Comparison to Fitnets

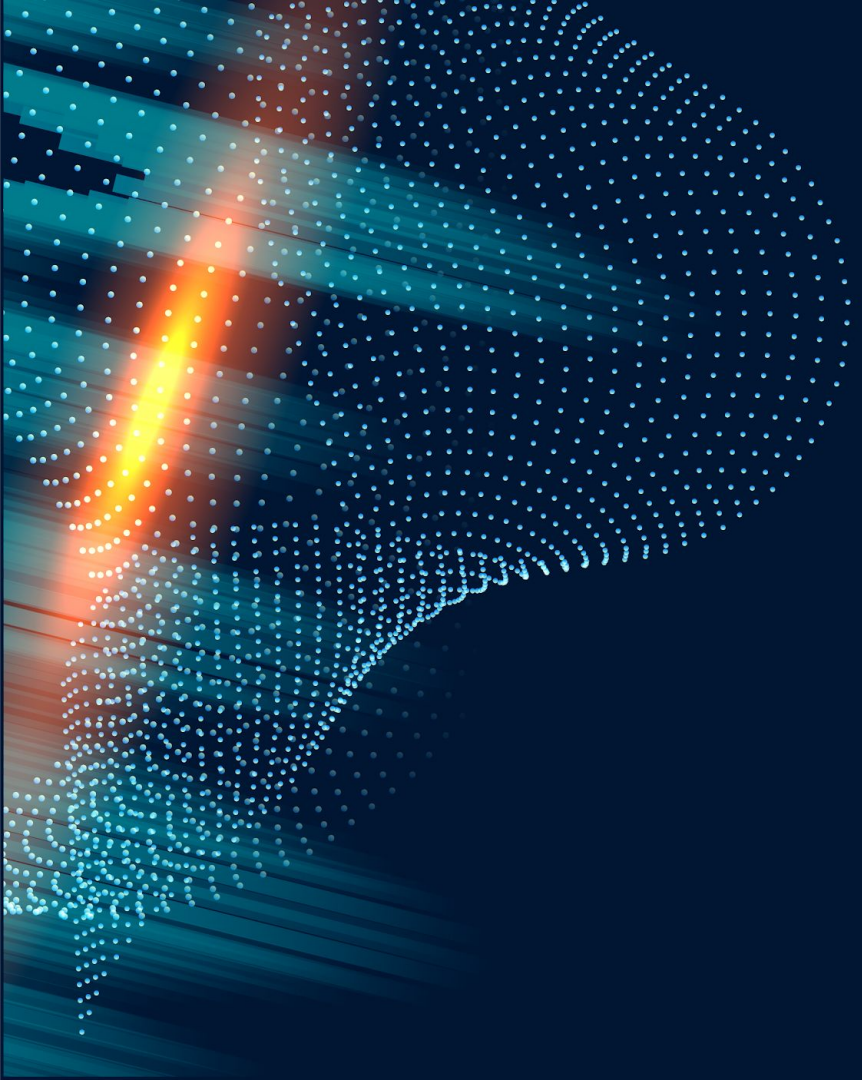
- Easy to train highway networks with numbers of parameters and operations comparable to those of fitnets in a single stage using SGD.
- higher accuracy on the test set.

Network	No. of Layers	No. of Parameters	Accuracy (in %)
Fitnet Results (reported by Romero et. al.[25])			
Teacher	5	~9M	90.18
Fitnet A	11	~250K	89.01
Fitnet B	19	~2.5M	91.61
Highway networks			
Highway A (Fitnet A)	11	~236K	89.18
Highway B (Fitnet B)	19	~2.3M	**92.46 (92.28±0.16)**
Highway C	32	~1.25M	91.20

Comparison to State-of-the-art Methods

- Performed experiments in the more common setting of global contrast normalization, small translations and mirroring of images
- Replaced the fully connected layer used in the networks in the previous section with a convolutional layer with a receptive field of size one and a global average pooling layer
- Quite possible to obtain much better results with better architectures/ hyperparameters

Network	CIFAR-10 Accuracy (in %)	CIFAR-100 Accuracy (in %)
Maxout [20]	90.62	61.42
dasNet [36]	90.78	66.22
NiN [35]	91.19	64.32
DSN [24]	92.03	65.43
All-CNN [37]	**92.75**	66.29
Highway Network	92.40 (92.31±0.12)	**67.76 (67.61±0.15)**



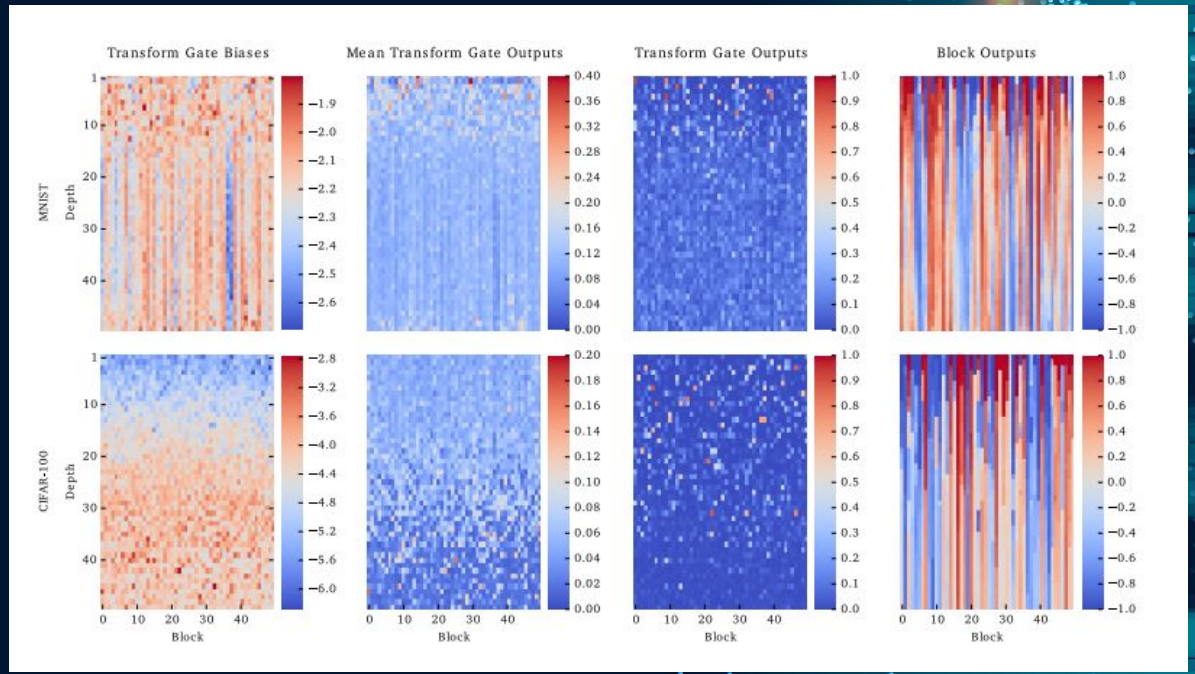
ANALYSIS

VISUALIZATION OF BEST 50 HIDDEN-LAYER HIGHWAY NETWORKS

For a single random training sample

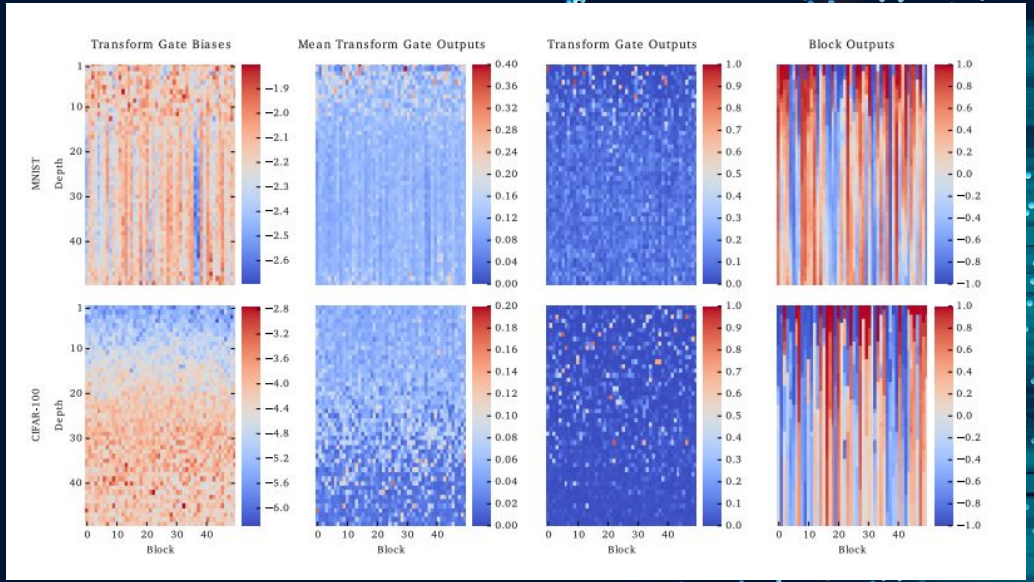
Trained on MNIST

Trained on CIFAR-100



- Most biases decreased further during training.

- For the CIFAR-100 network, the biases increase with depth forming a gradient which is inversely correlated with the average activity of the transform gates (second column).



- Strong negative biases at low depths are not used to shut down the gates, but to make them more selective.

- Most of the outputs stay constant over many layers forming a pattern of stripes. Most of the change in outputs happens in the early layers (≈ 15 for MNIST and ≈ 40 for CIFAR-100).

Routing of Information

- One possible advantage of the highway architecture is that the network can learn to dynamically adjust the routing of the information based on the current input.

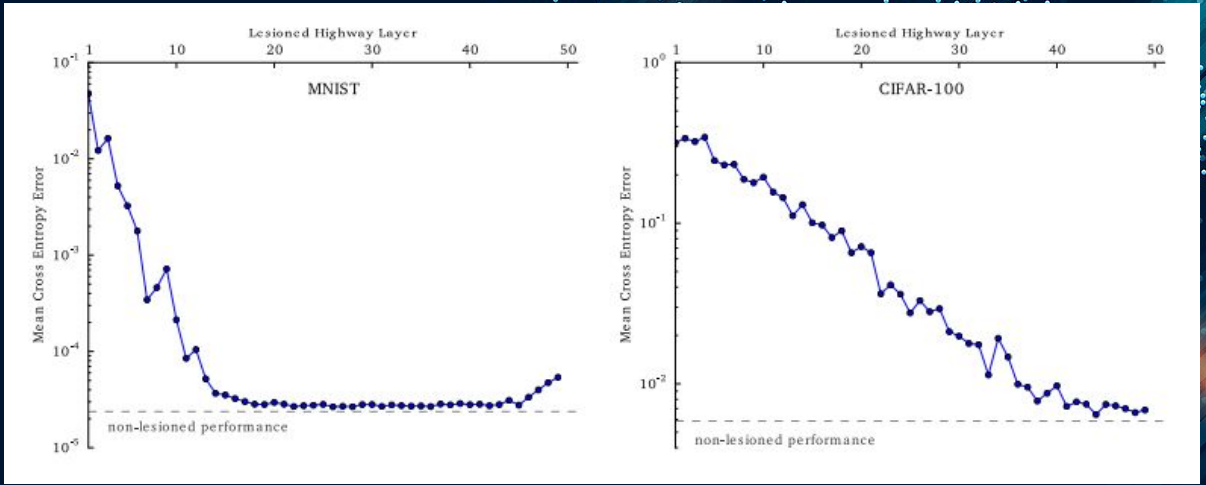
- In the mean transform gate activity (second column) and the single example transform gate outputs (third column), most transform gates are active on average and show very selective activity for the single example.

- As a data-dependent routing mechanism, this implies that only a few blocks perform transformation but different blocks are utilized by different samples.

- For MNIST digits 0 and 7 substantial differences can be seen within the first 15 layers, while for CIFAR class numbers 0 and 1, the differences spread out over all layers. In both cases it is clear that the mean activity pattern differs between classes.

Layer Importance

- For MNIST (left), the error rises significantly if any one of the early layers is removed
- Layers 15 – 45 seem to have close to no effect on the final performance.

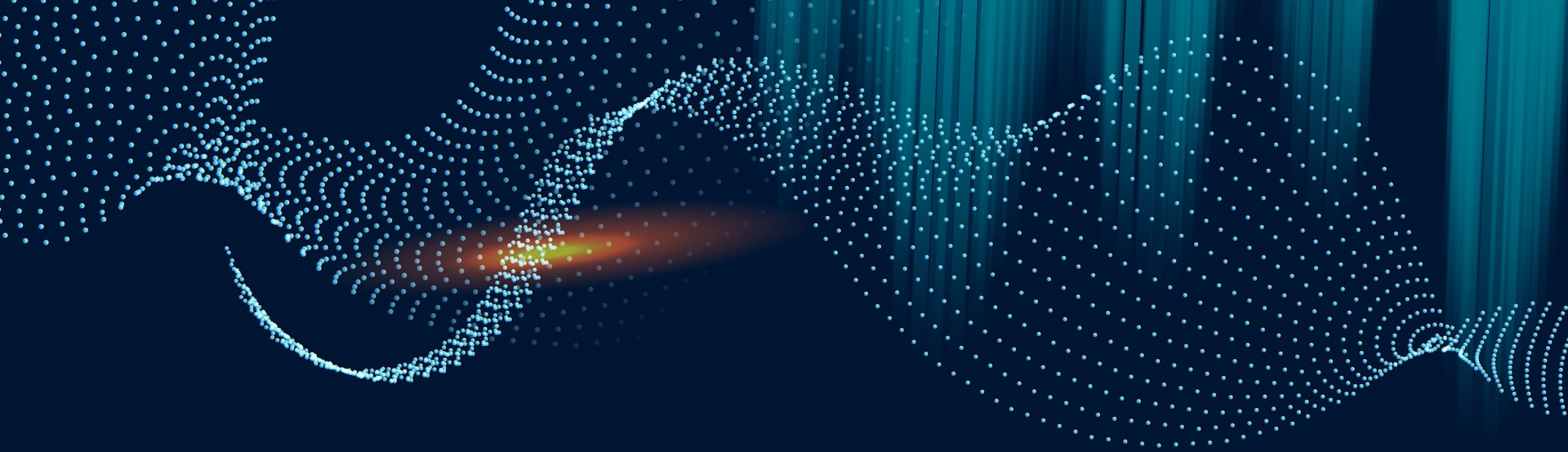


Resulting performance as a function of the lesioned layer

- About 60% of the layers don't learn to contribute to the final result, likely because MNIST is a simple dataset that doesn't require much depth.
- CIFAR-100 dataset (right), performance degrades noticeably when removing any of the first ≈40 layers. This suggests that for complex problems a highway network can learn to utilize all of its layers

SUMMARY

- Very deep highway networks can directly be trained with simple gradient descent methods due to their specific architecture. This property does not rely on specific non-linear transformations, which may be complex convolutional or recurrent transforms.
- A possible objection is that many layers might remain unused if the transform gates stay closed
- Deep and narrow highway networks can match/exceed the accuracy of wide and shallow maxout networks
- We can exploit the structure of highways to directly evaluate the contribution of each layer
- Highway networks allow us to examine how much computation depth is needed for a given problem, which can not be easily done with plain networks



THANK YOU

Q&A