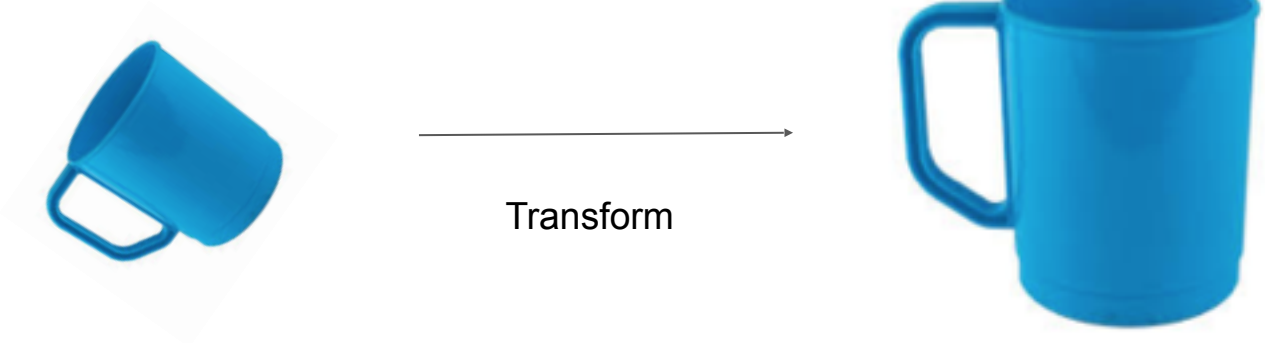
Spatial Transformer Network

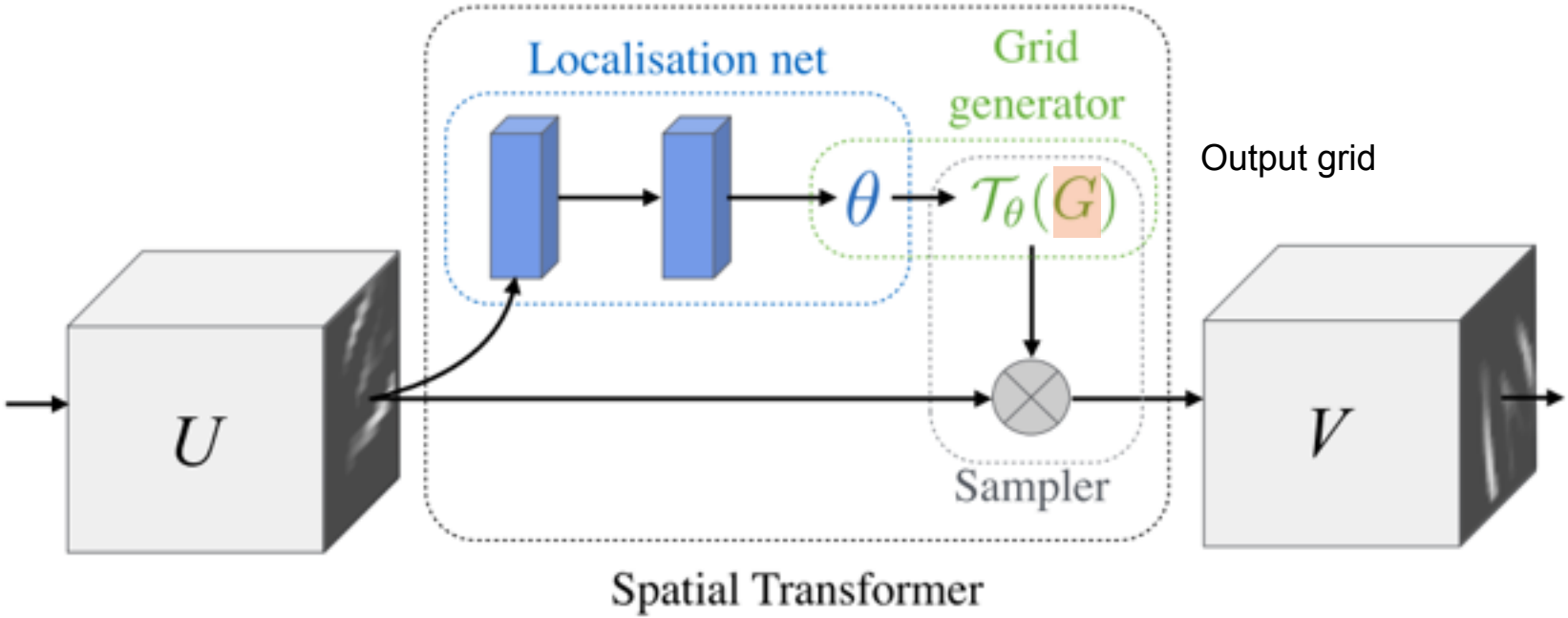
Jiaxuan Wang
4/16/2018

Motivation

Learning invariances in image data for robust prediction



How does it work?



Jaderberg et. al. (2015)

Experiment 1: Rotation + Translation + Scaling

Random Rotation: [-45, 45] degrees

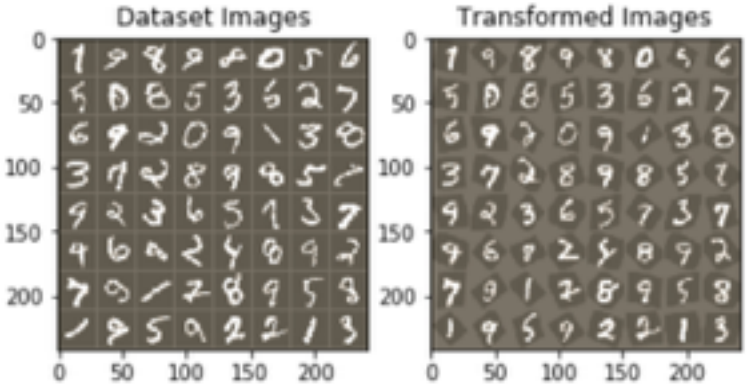
Random Translation: [-0.1, 0.1] independently for each x, y axis

Random Scaling: [0.7, 1.3] along both directions

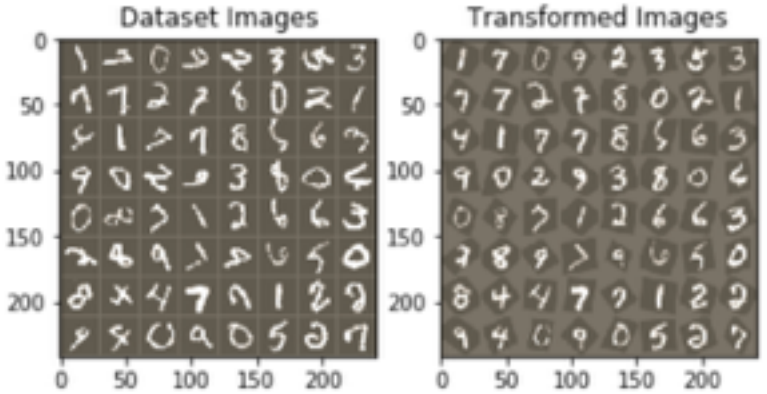
Baseline: CNN with comparable number of neurons

Test: Generalization of STN on rotation, translation, and scaling beyond training data augmentation

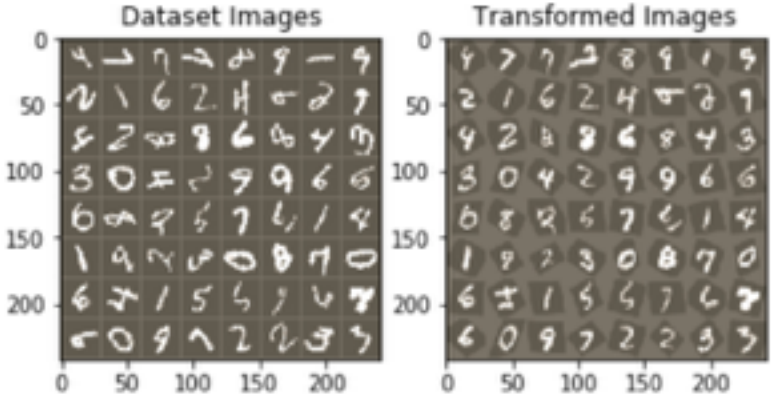
45 degree



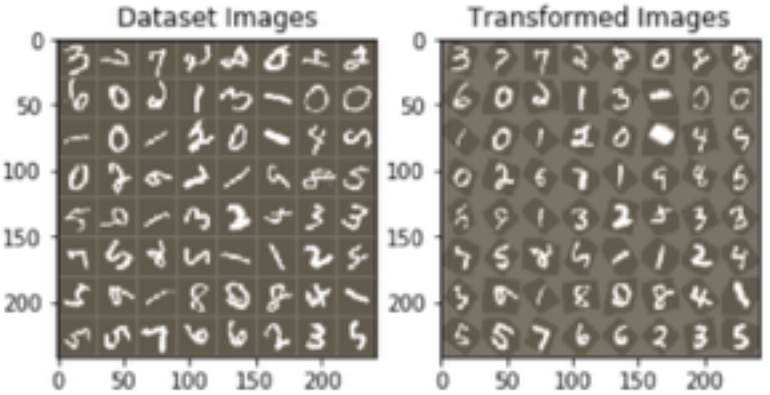
60 degree

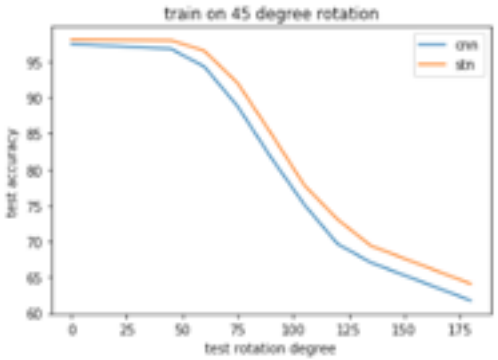


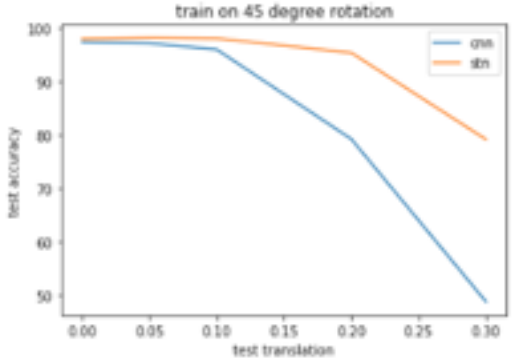
75 degree

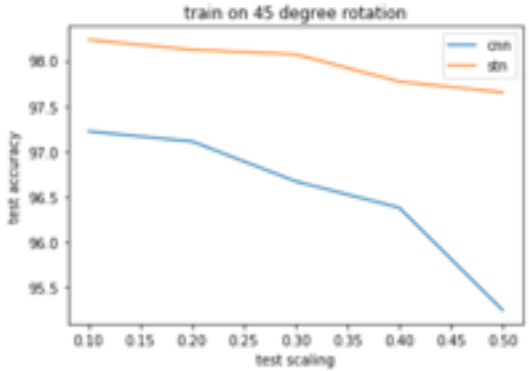


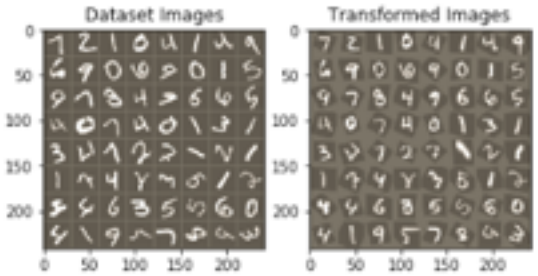
90 degree

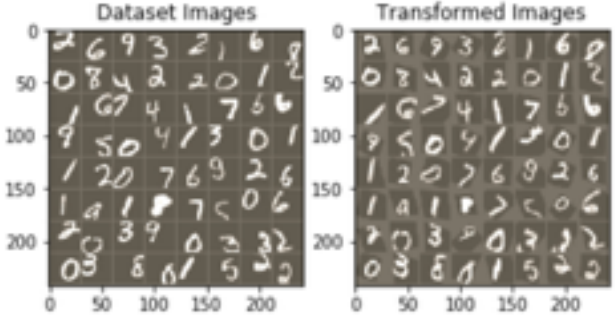


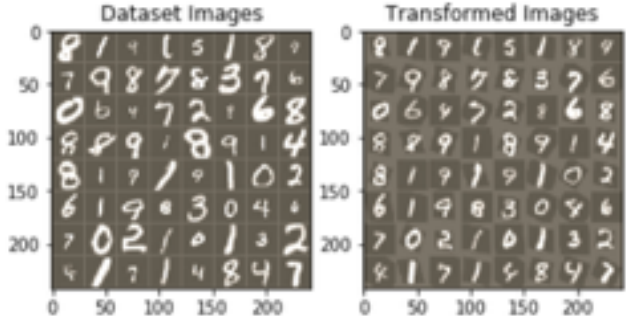

train on 45 degree rotation
cnn
stn
test accuracy
test rotation degree


train on 45 degree rotation
cnn
stn
test accuracy
test translation

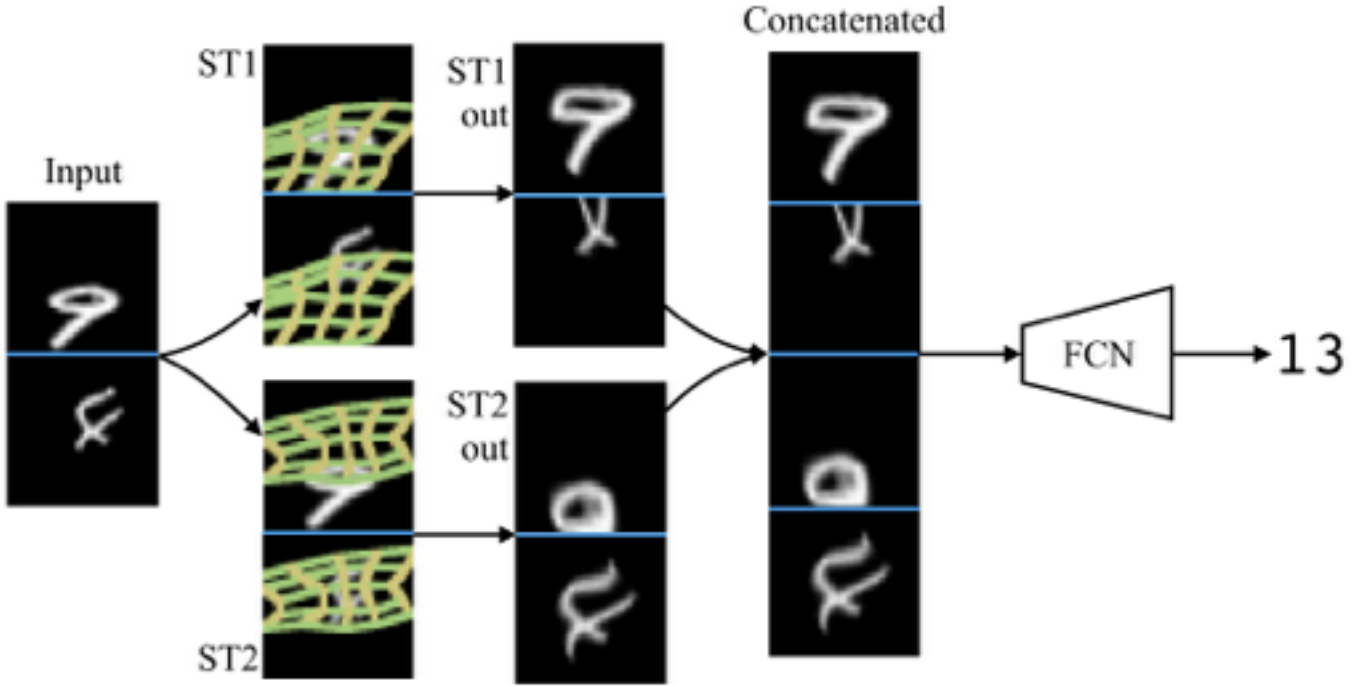

train on 45 degree rotation
cnn
stn
test accuracy
test scaling


Dataset Images
Transformed Images


Dataset Images
Transformed Images

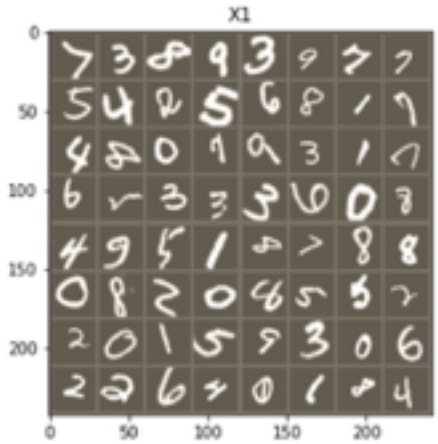

Dataset Images
Transformed Images

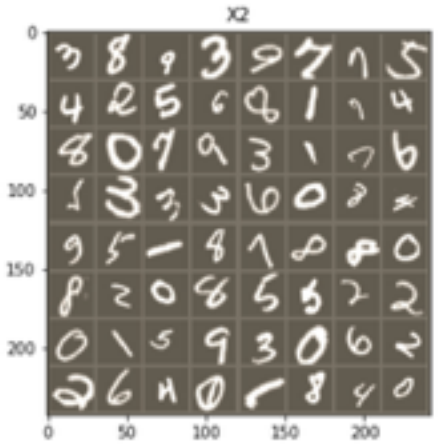
Experiment 2: MNIST addition

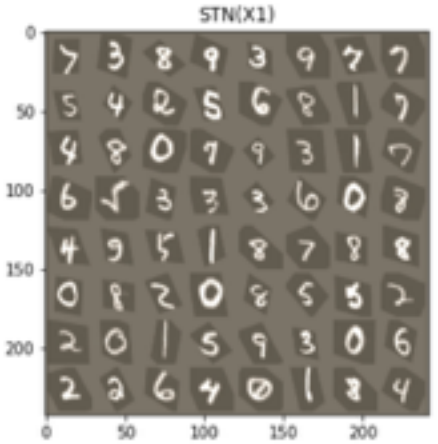


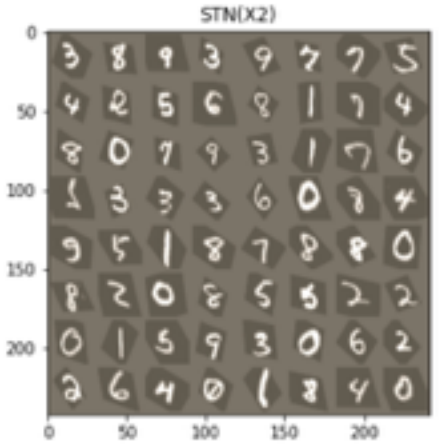
Jaderberg et. al. (2015)

STN performance: 93.47%
CNN performance: 85.30%





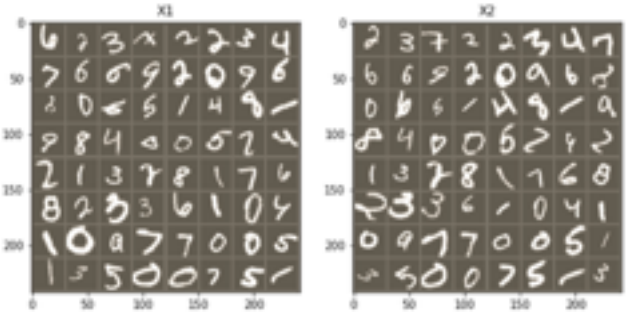




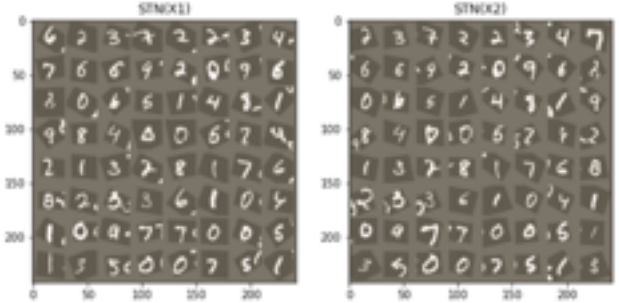
Experiment 3: MNIST addition on same image

Sensitivity to initialization

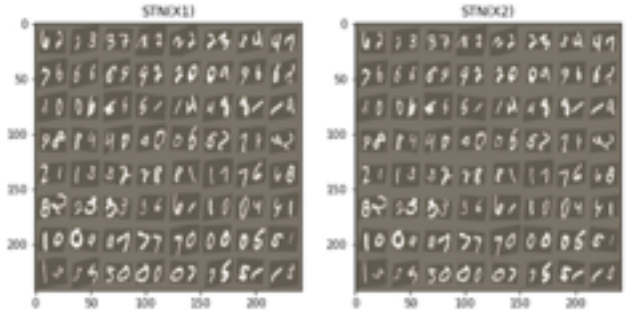
Data, CNN baseline 85% accuracy



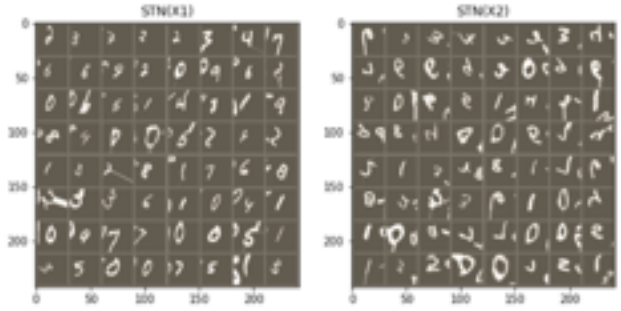
LR 92% accuracy



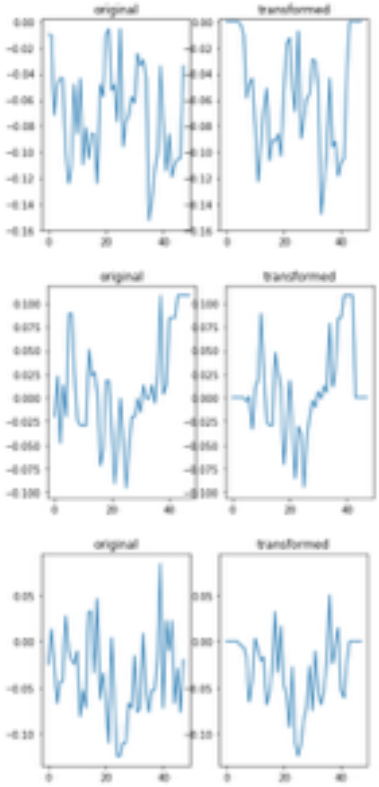
Identity 90% accuracy



Random 69% accuracy



Experiment 4: STN on time series data



Potential Issues:

a) Interpolation smoothes the difference
b) Currently only learns to shrink the time series but never expand
c) Hard to know what canonical means in this situation

	AUC
LSTM[1]	85.40%
STN	84.96%
CNN	83.48%

[1]Harutyunyan, Hrayr, et al. "Multitask Learning and Benchmarking with Clinical Time Series Data." arXiv preprint arXiv:1703.07771 (2017).

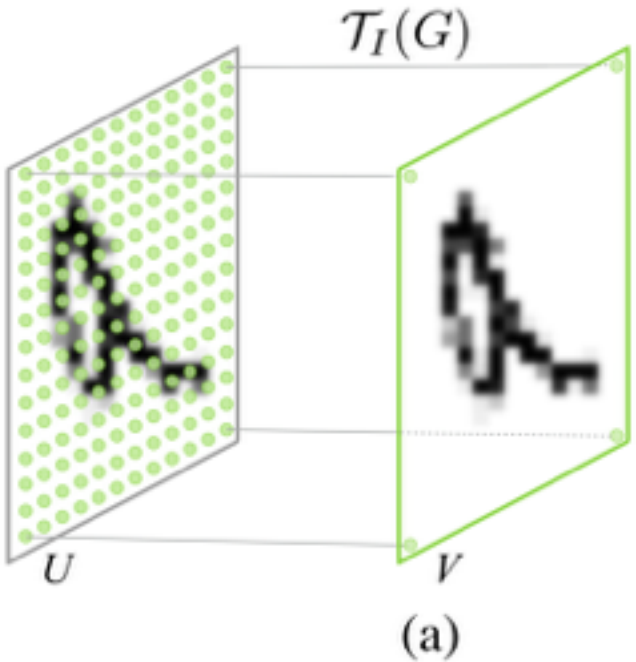
Work with Jeeheh Oh

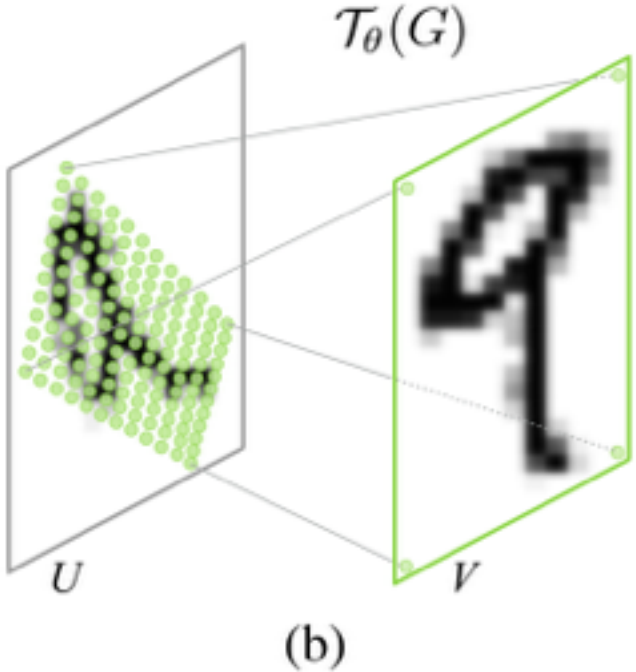
Lesson Learned

- Specify network parameters is key to reproduction
- Tradeoff between human learning and machine learning

Questions

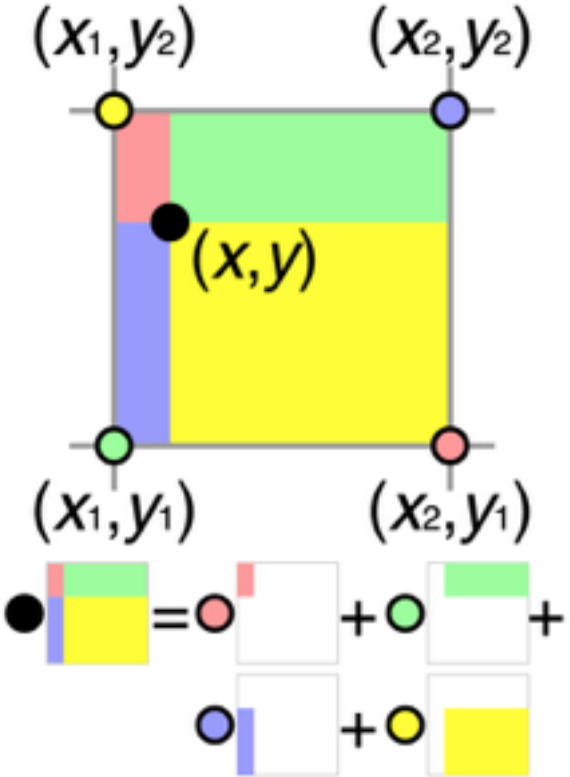
How does it work?





Jaderberg et. al. (2015)

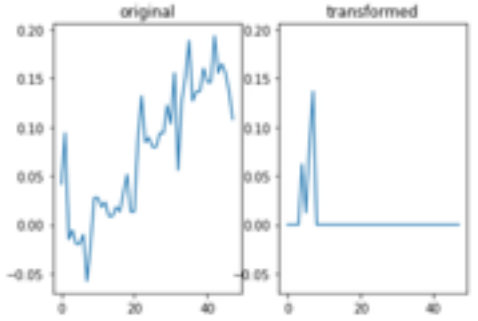
How to sample grid: bilinear interpolation

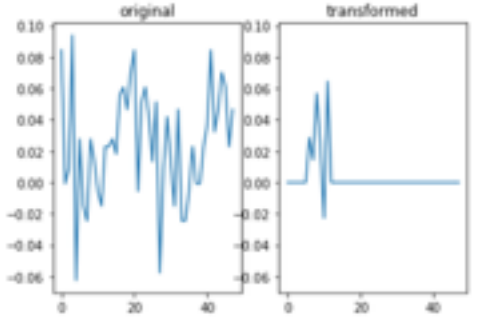


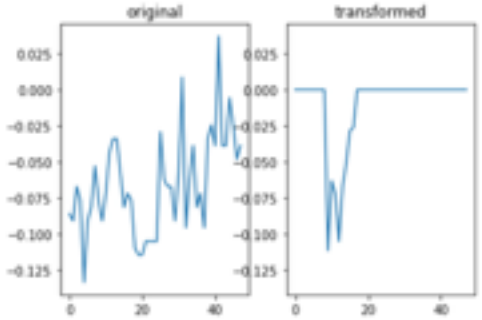
$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

In fact, you just need to sample nearest 4 points

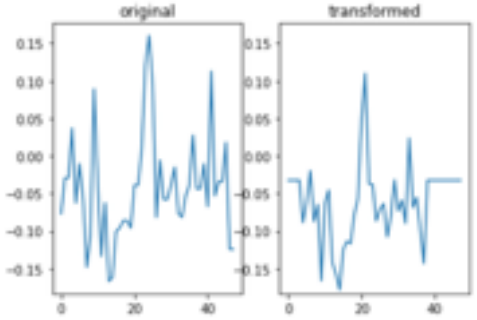
$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

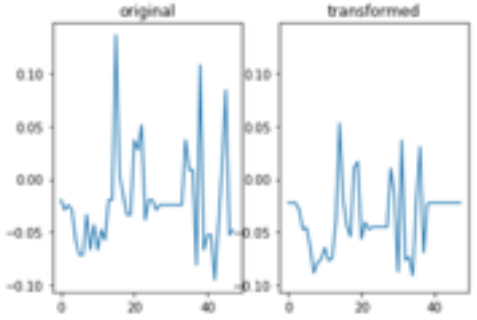


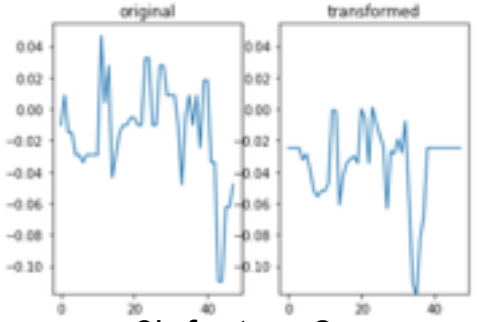




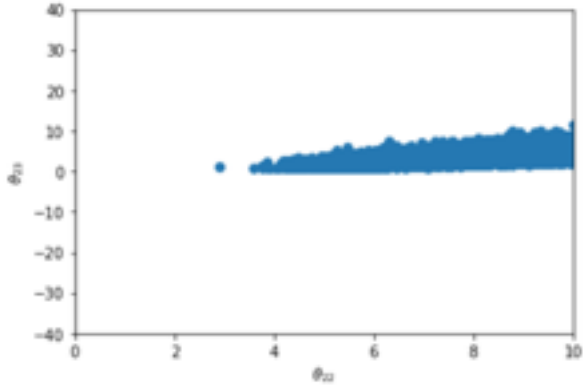
7b feature 2







8b feature 2


40
30
20
10
0
−10
−20
−30
−40
0
2
4
6
8
10
θ_{21}
θ_{22}