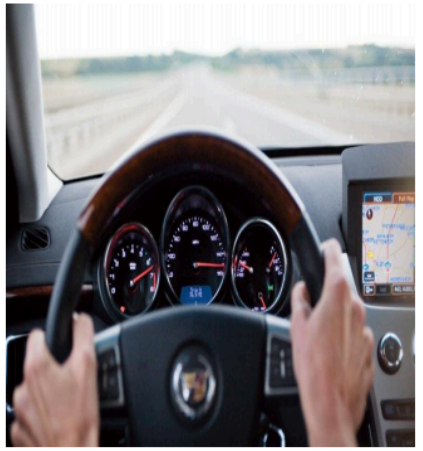
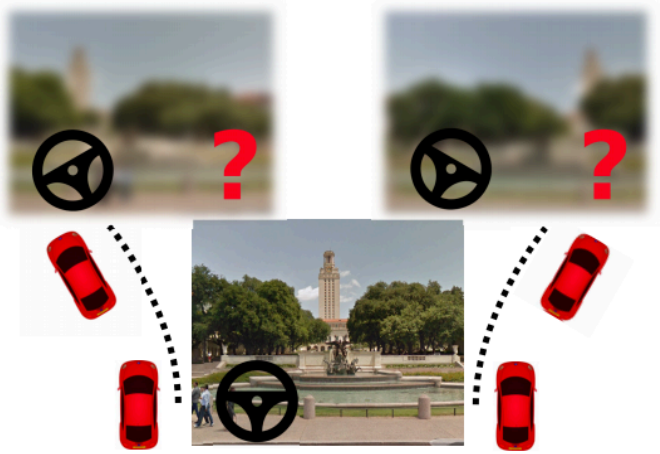
Learning image representations tied to ego-motion

Dinesh Jayaraman and Kristen Grauman

(experiment presentation)

Presented by Ruohan Gao
UTCS CS381V
Visual Recognition
2016 Spring

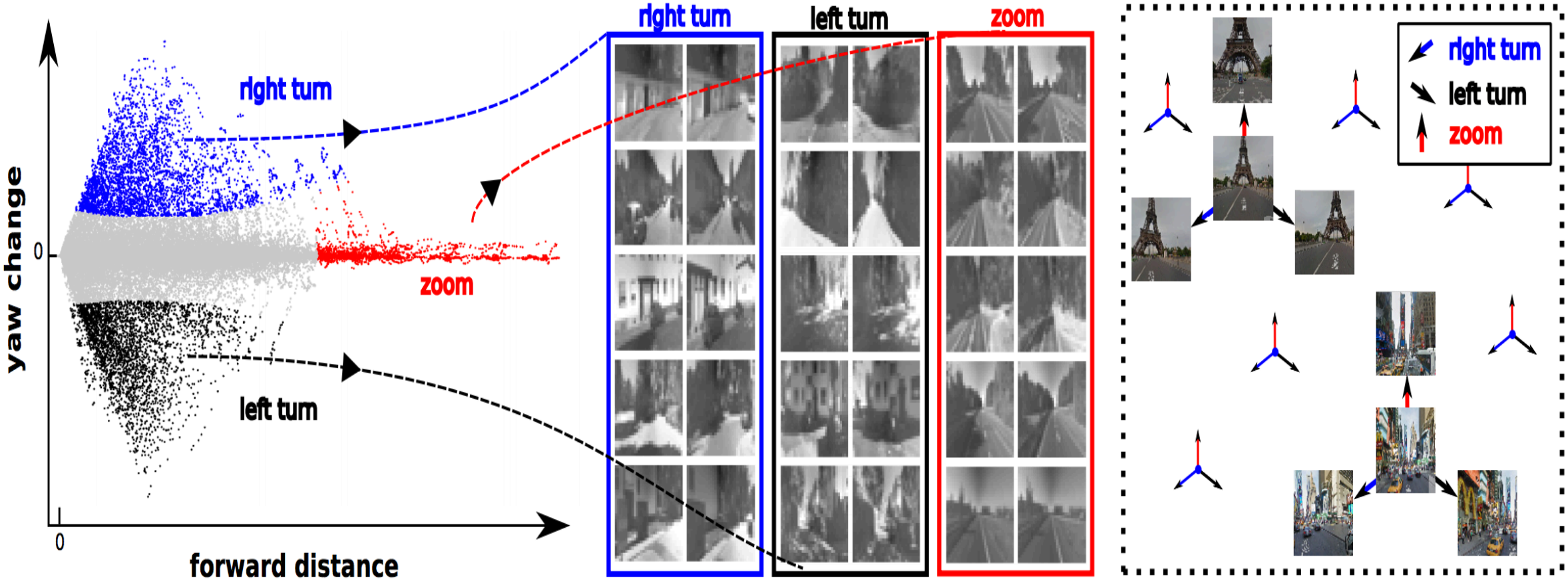




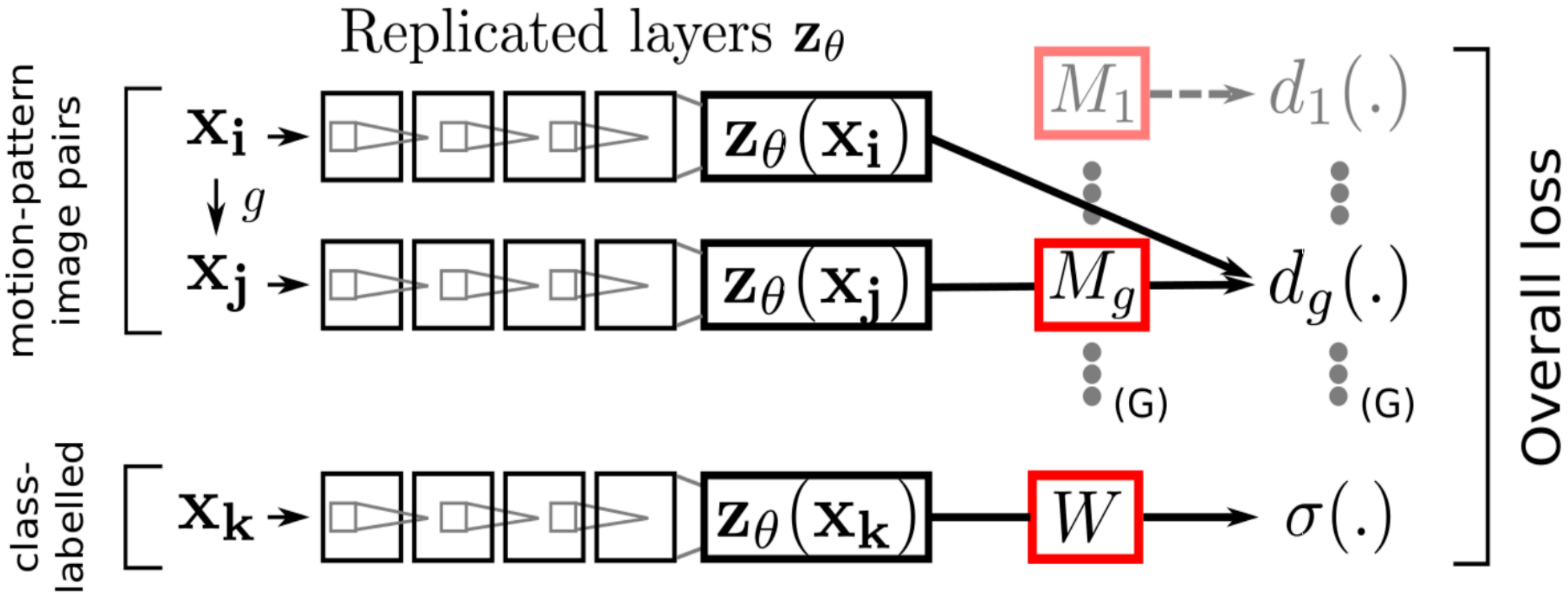
Outline

- Feature Embedding using Siamese Network
- Visualization of Learnt Features
- Fine-tuning on New Task

Feature Embedding using Siamese Network



Feature Embedding using Siamese Network

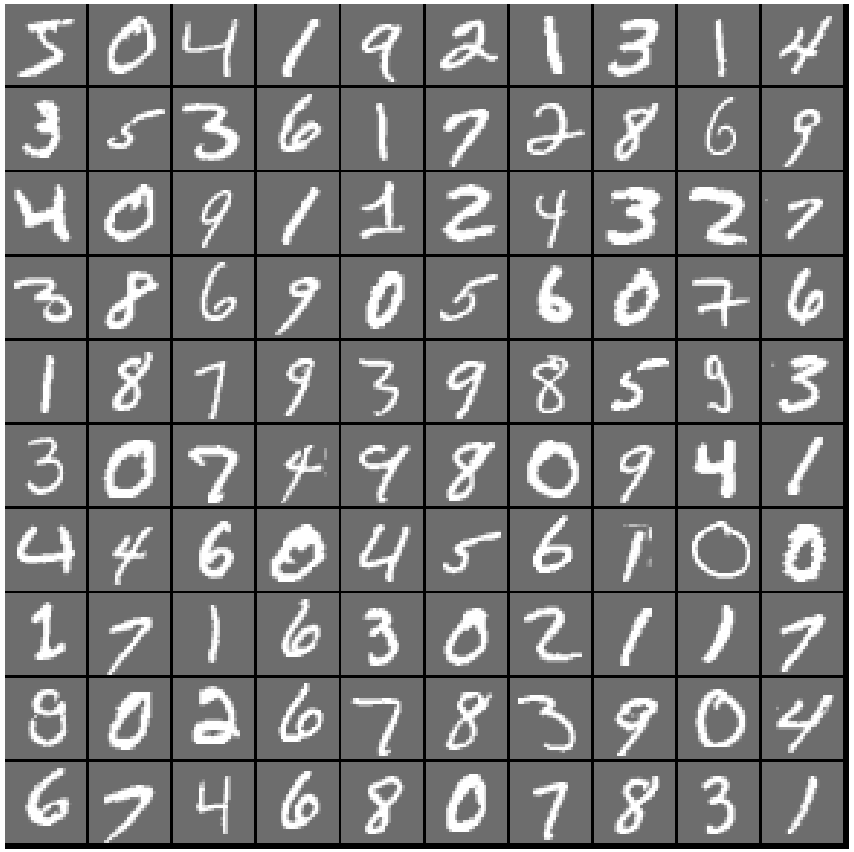


Feature Embedding using Siamese Network

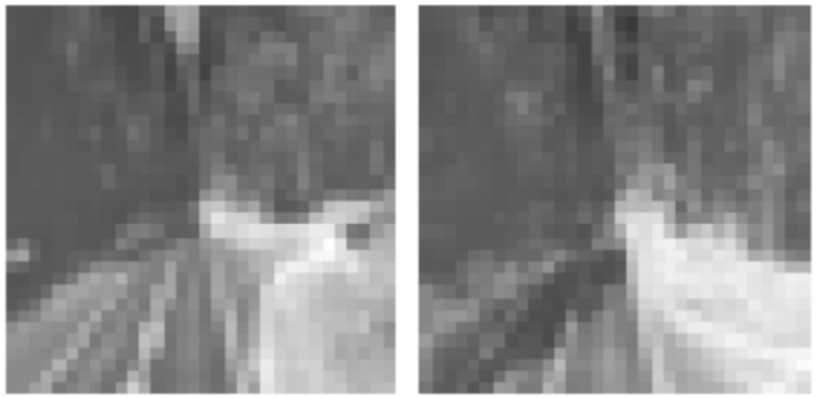
Contrastive Loss:

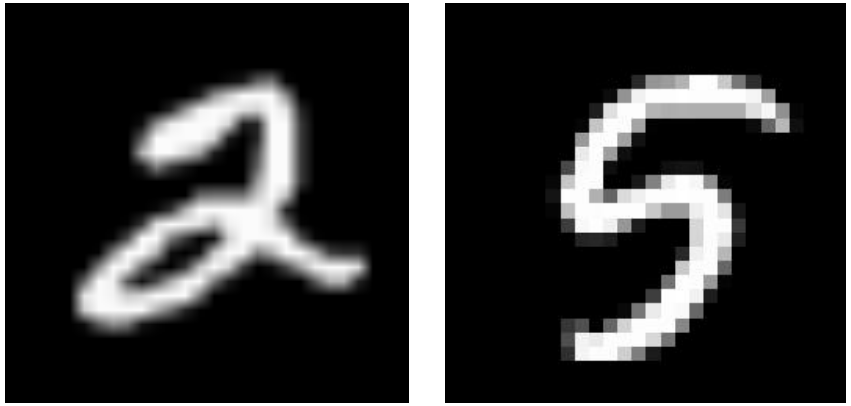
$$d_g(\boldsymbol{a}, \boldsymbol{b}, c) = \mathbb{1}(c = g)d(\boldsymbol{a}, \boldsymbol{b}) + \mathbb{1}(c \neq g)\max(\delta - d(\boldsymbol{a}, \boldsymbol{b}), 0)$$

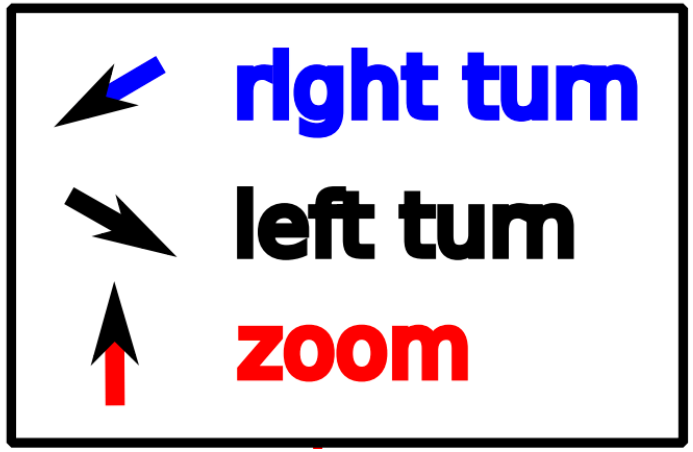
Toy Example of MNIST Dataset

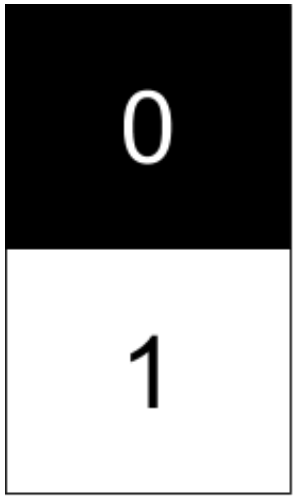


Toy Example of MNIST Dataset



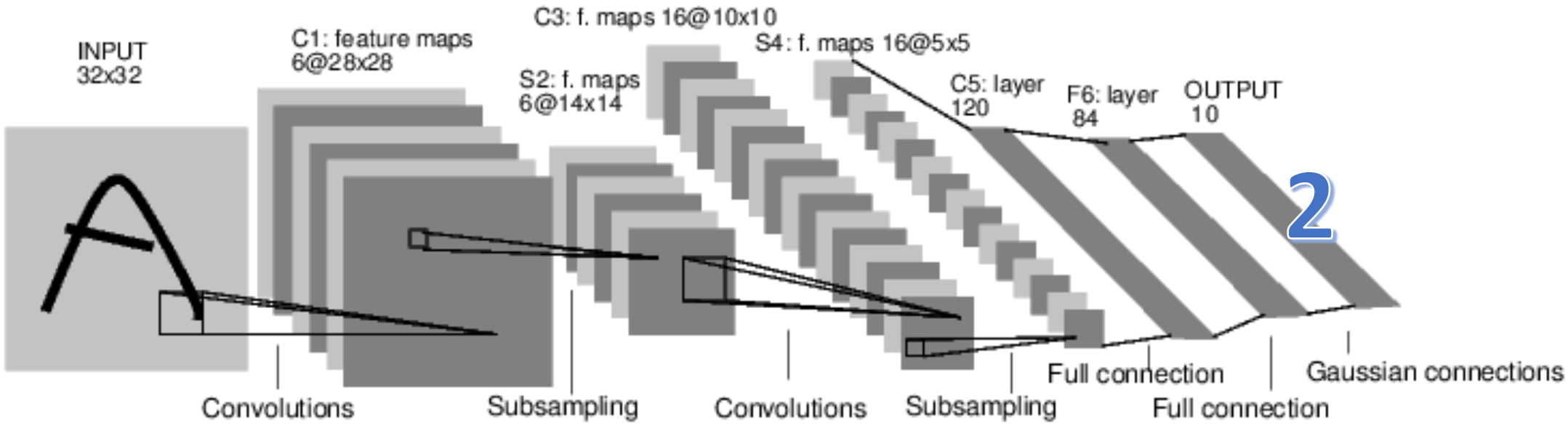




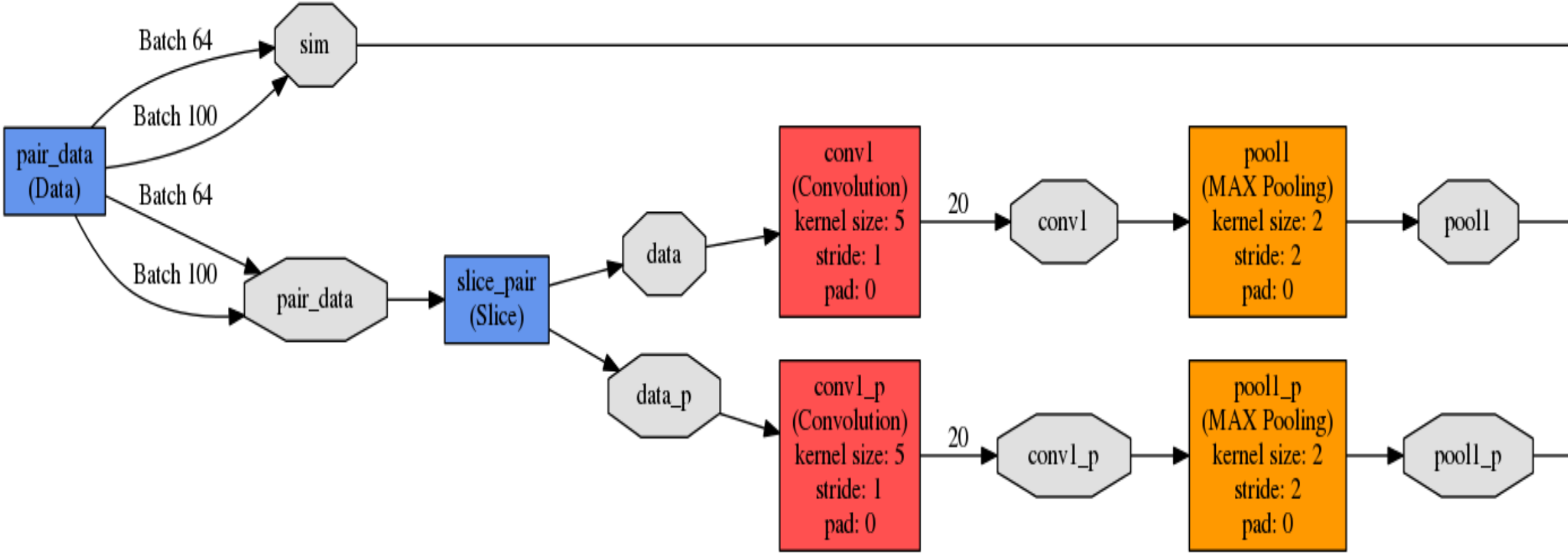


Toy Example of MNIST Dataset

LeNet5

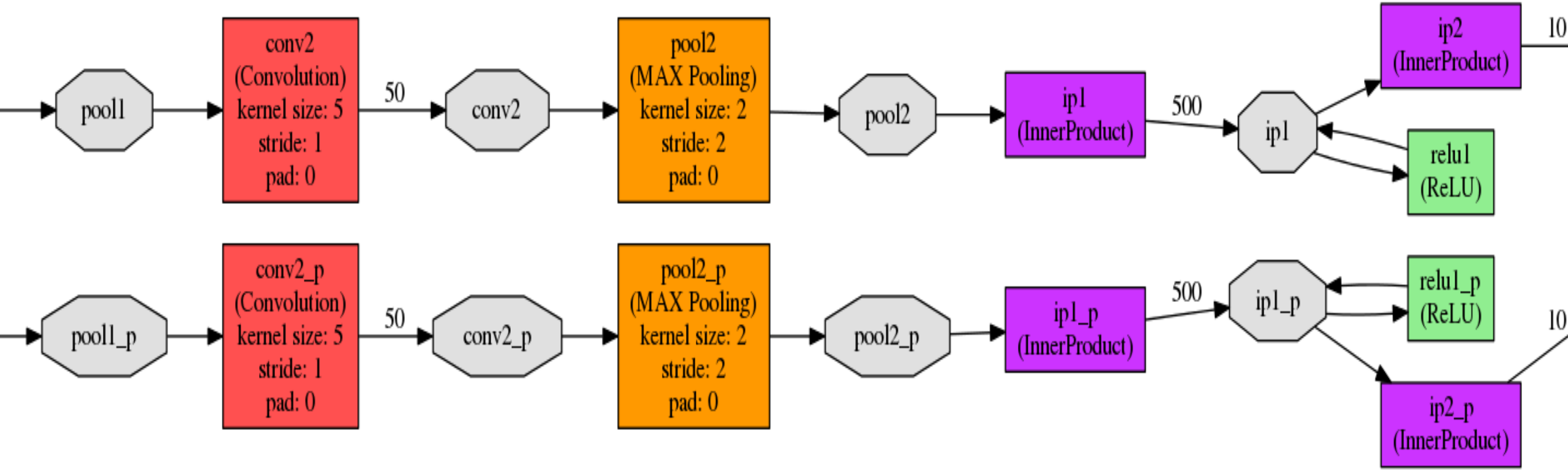


Toy Example of MNIST Dataset



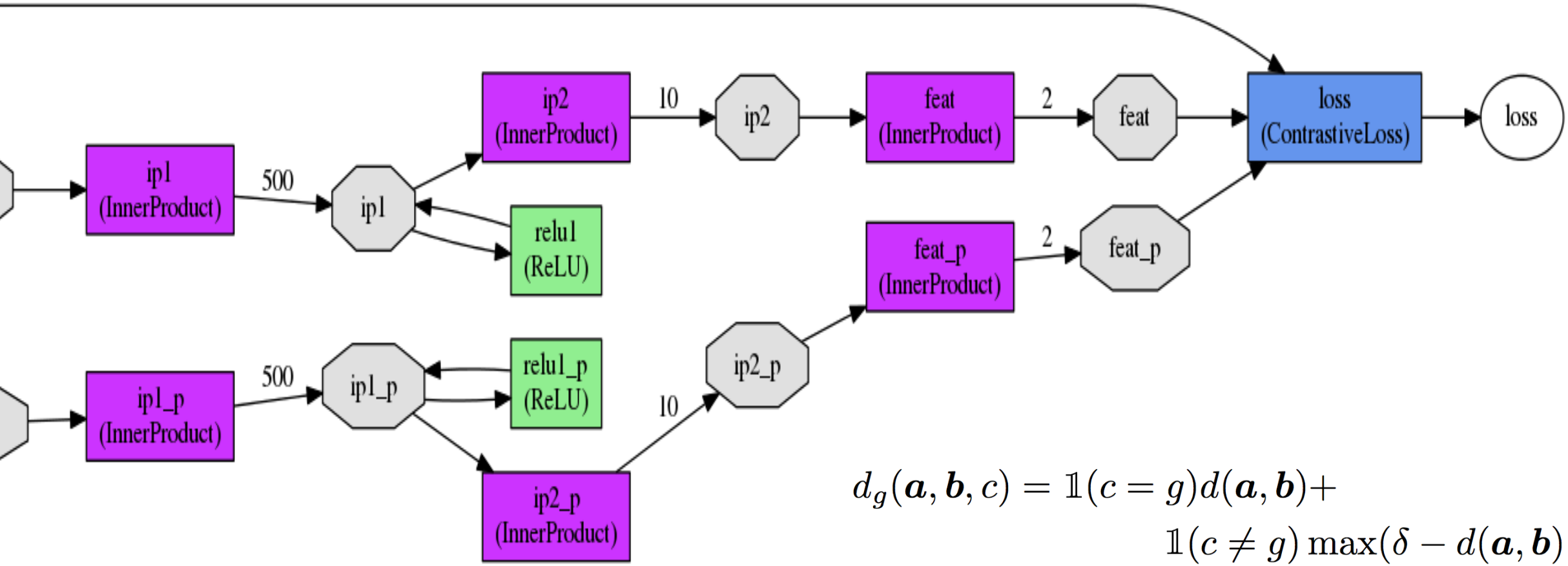
Generated using Caffe tool: draw_net.py

Toy Example of MNIST Dataset



Generated using Caffe tool: draw_net.py

Toy Example of MNIST Dataset

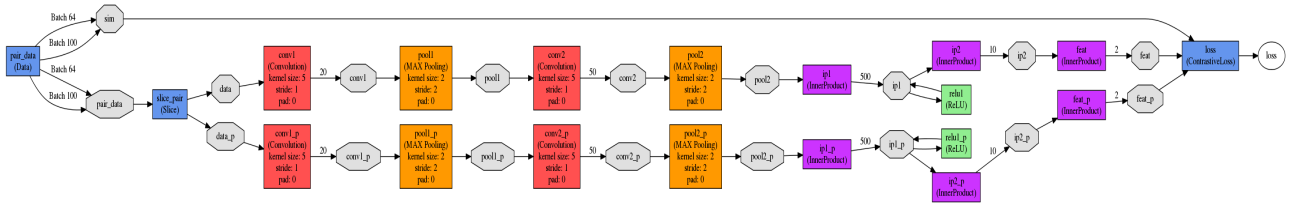


$$d_g(\boldsymbol{a}, \boldsymbol{b}, c) = \mathbb{1}(c = g)d(\boldsymbol{a}, \boldsymbol{b}) + \mathbb{1}(c \neq g)\max(\delta - d(\boldsymbol{a}, \boldsymbol{b}), 0)$$

Generated using Caffe tool: draw_net.py

Toy Example of MNIST Dataset

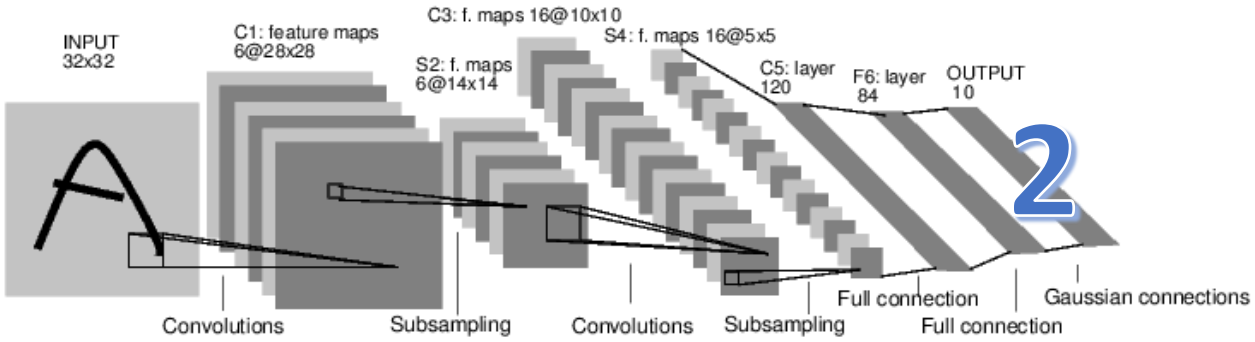
- Settings of Training Phase:
  - Base Learning Rate: 0.01
  - Momentum: 0.9
  - Weight Decay: 0
  - Learning Rate Policy: “inv”
  - gamma: 0.0001
  - power: 0.75
  - Training Data: 10,000 pairs
  - Batch Size: 64
  - Number of Iterations: **20,000**



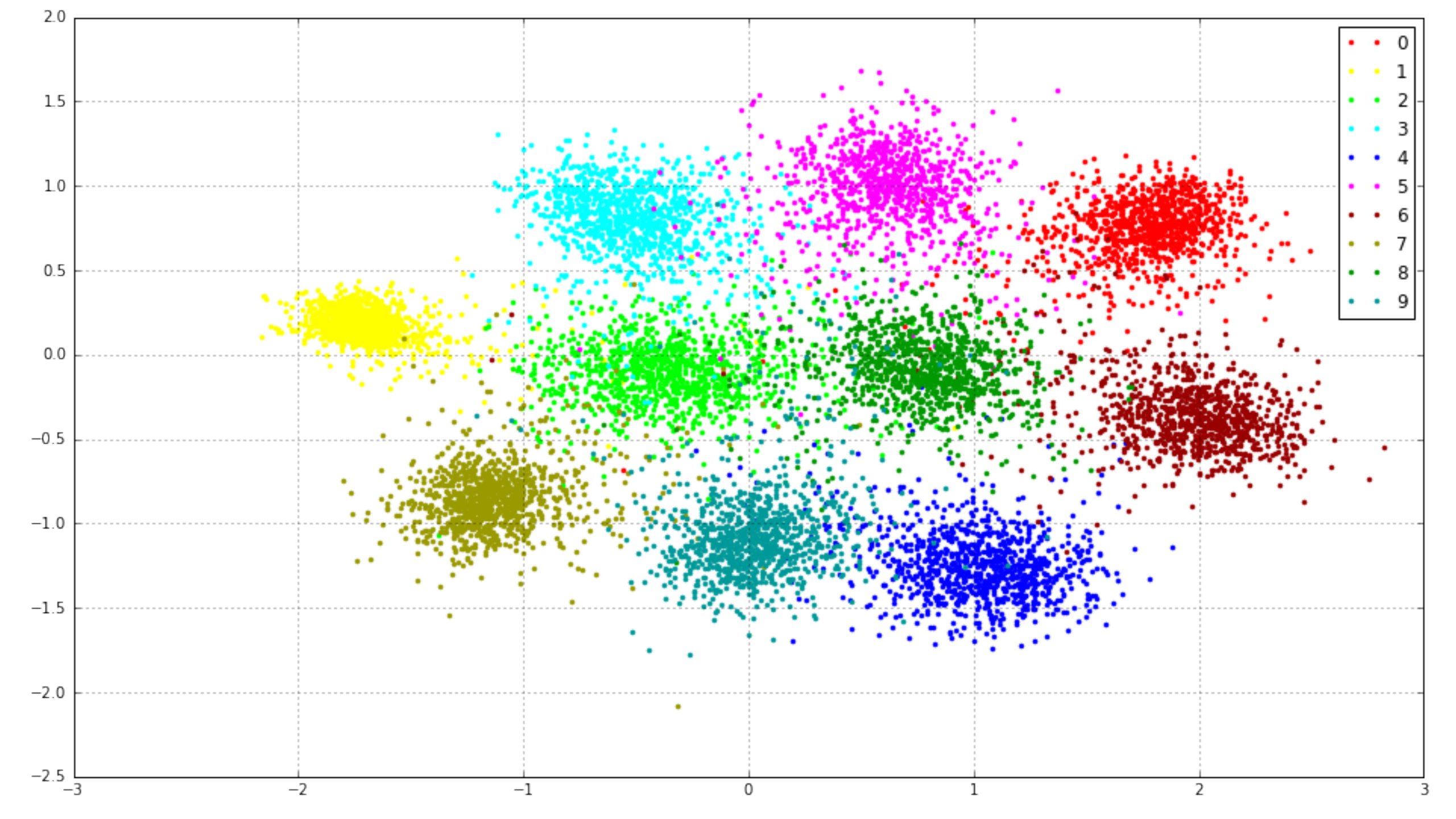
Modified from Caffe official Tutorial: http://caffe.berkeleyvision.org/

Toy Example of MNIST Dataset

- Feature Embedding Setting:
  - Number of Testing Digits: 10,000
  - mnist_siamese_iter_20000.caffemodel
  - Embed in two-dimensional feature space

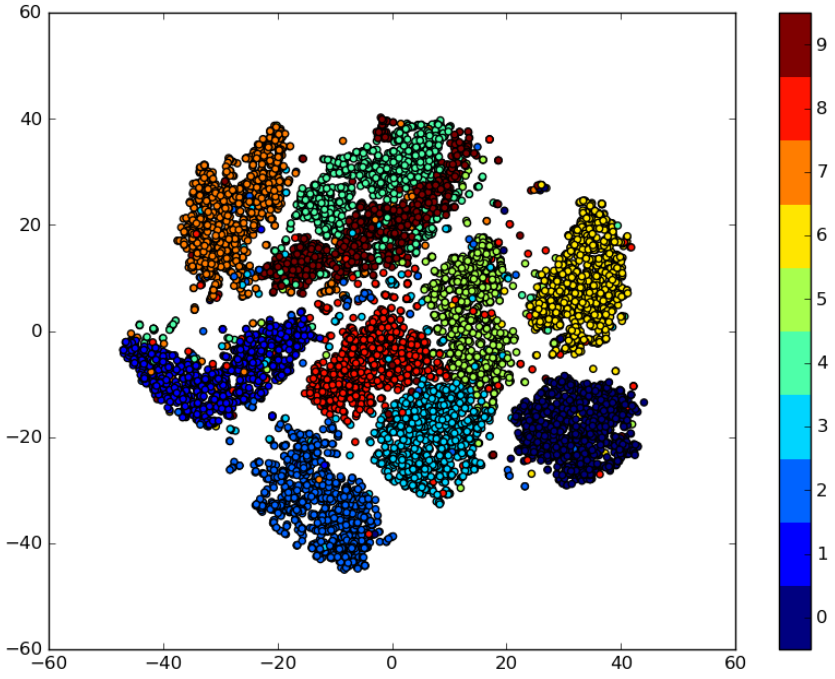


Modified from Caffe official Tutorial: http://caffe.berkeleyvision.org/


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3
−2
−1
0
1
2
3
0
1
2
3
4
5
6
7
8
9

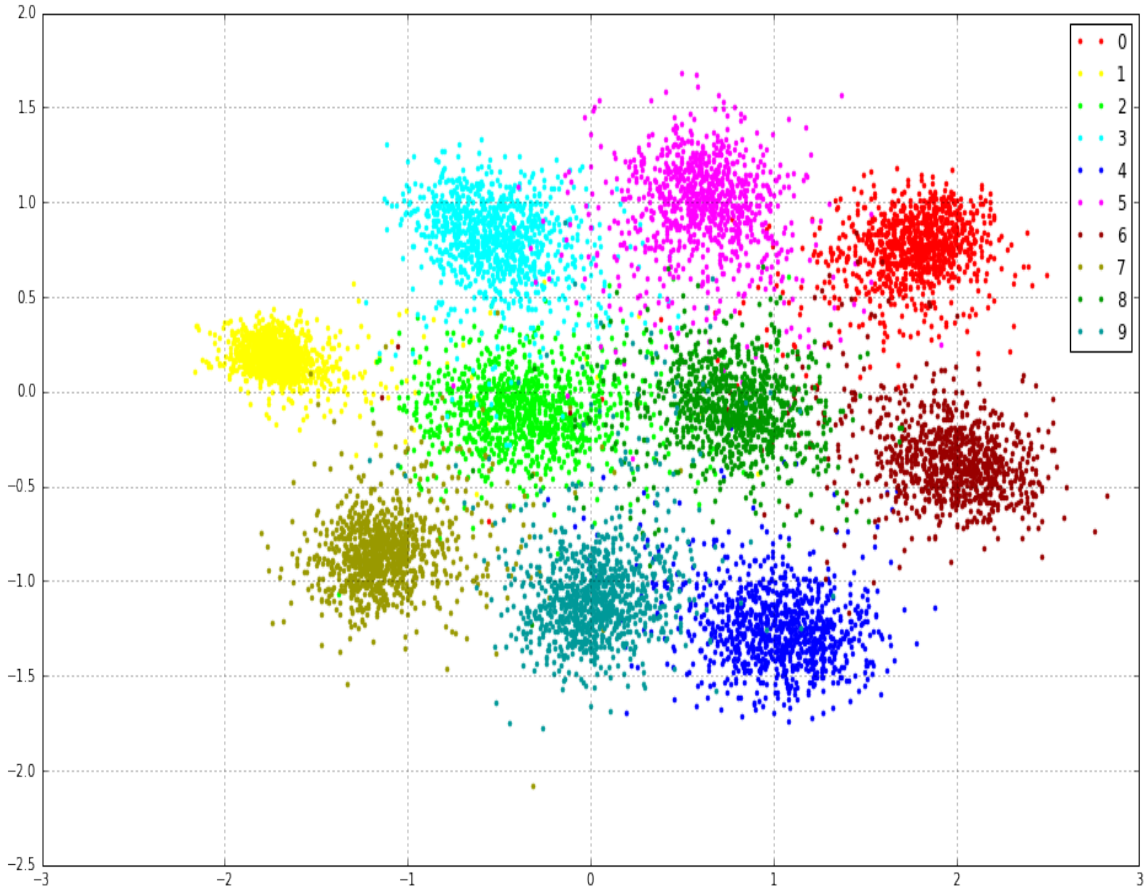
Toy Example of MNIST Dataset

- To Compare -> a dimensionality reduction algorithm: t-SNE (t-distributed stochastic neighbor embedding)
- Reduce the number of dimensions to two

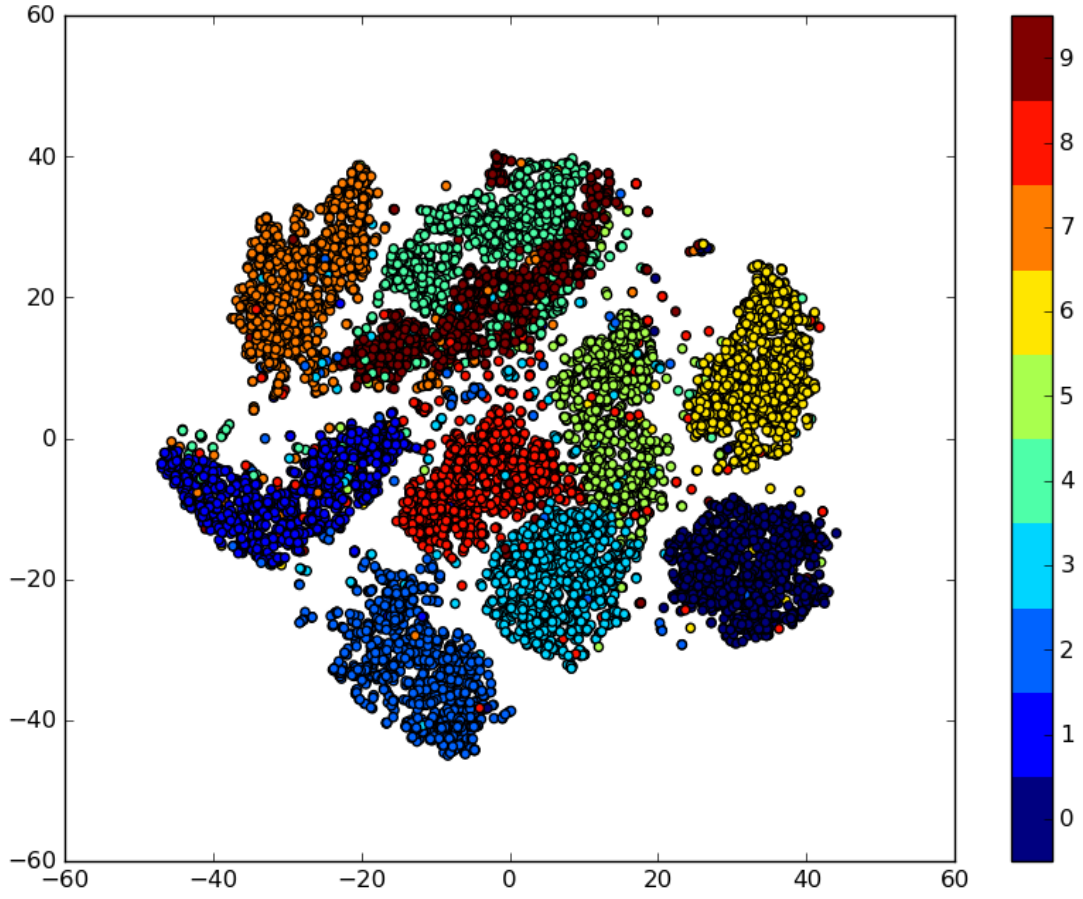


Ref:http://lvdmaaten.github.io/tsne/

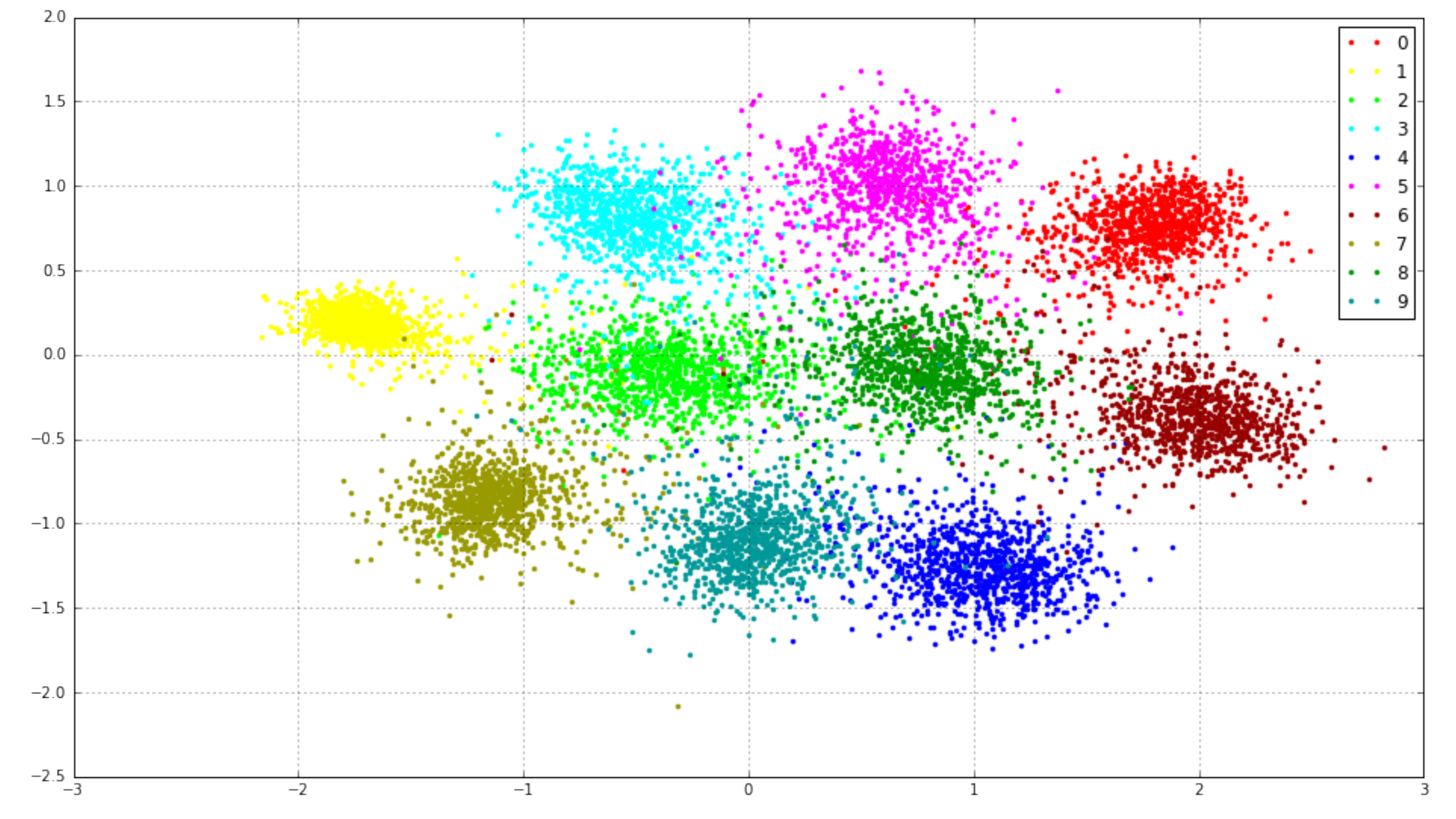
Toy Example of MNIST Dataset



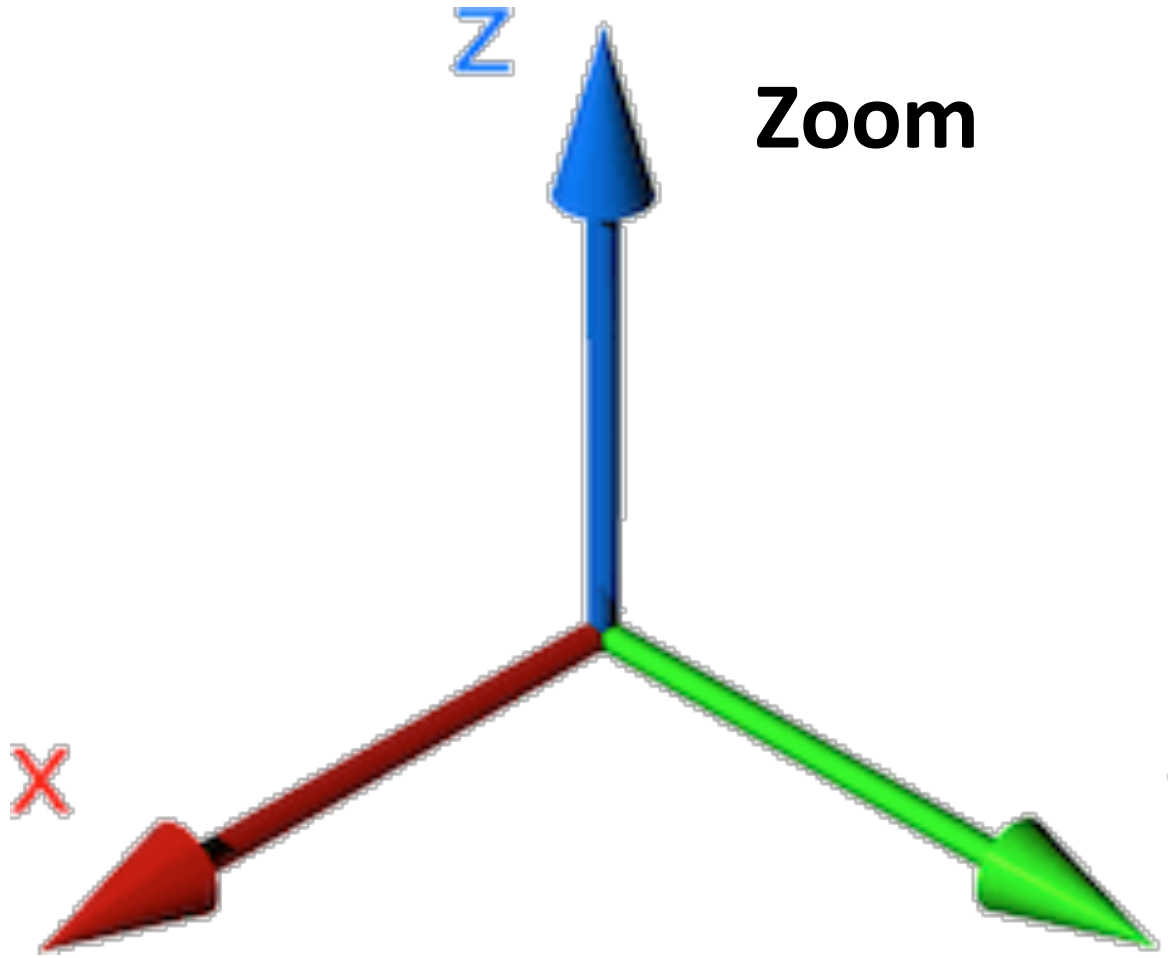
Siamese Embedding



t-SNE Embedding


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3
−2
−1
0
1
2
3
0
1
2
3
4
5
6
7
8
9

Invariance vs Equivariance

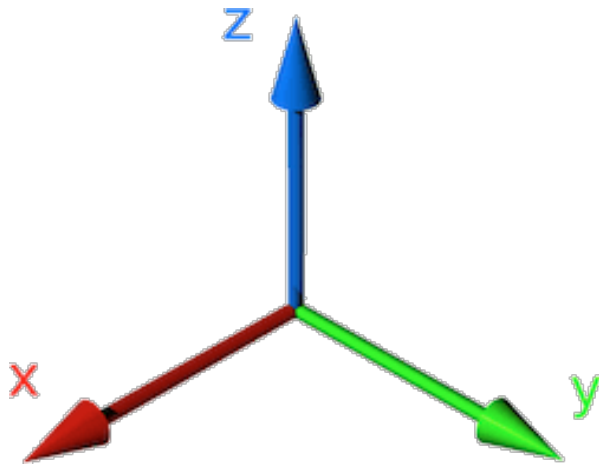


Zoom

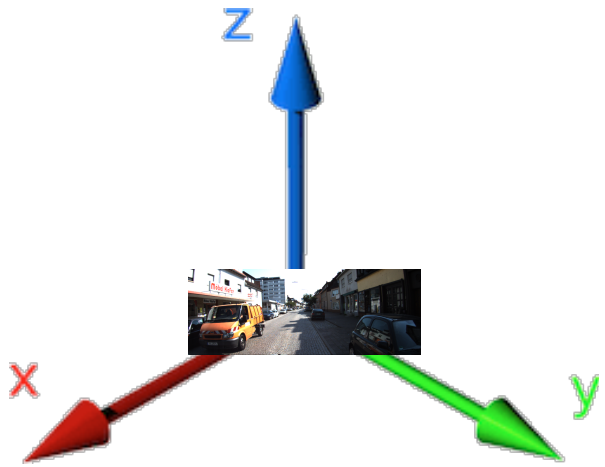
Right Turn

Left Turn

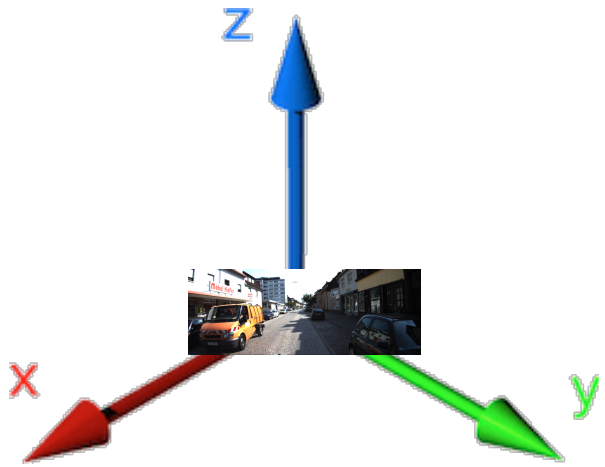
Invariance vs Equivariance



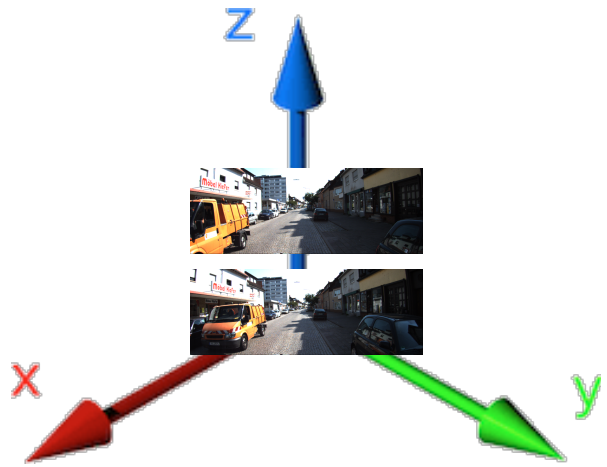
Invariance vs Equivariance



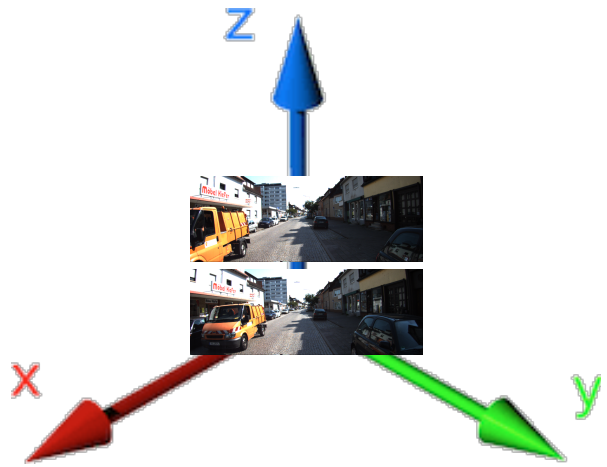
Invariance vs Equivariance



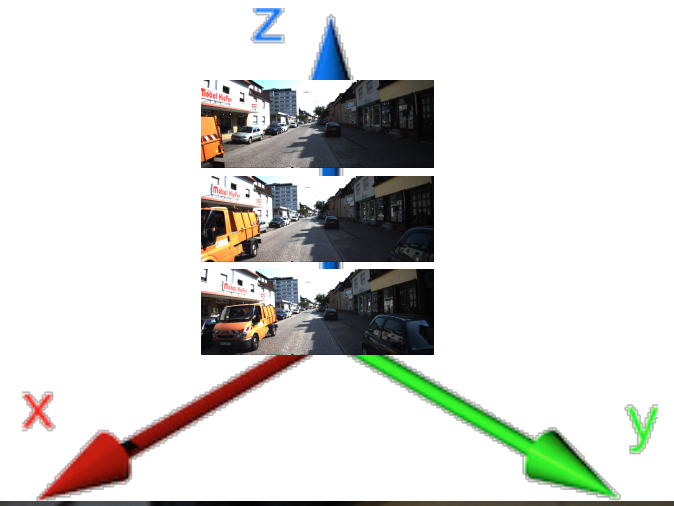
Invariance vs Equivariance



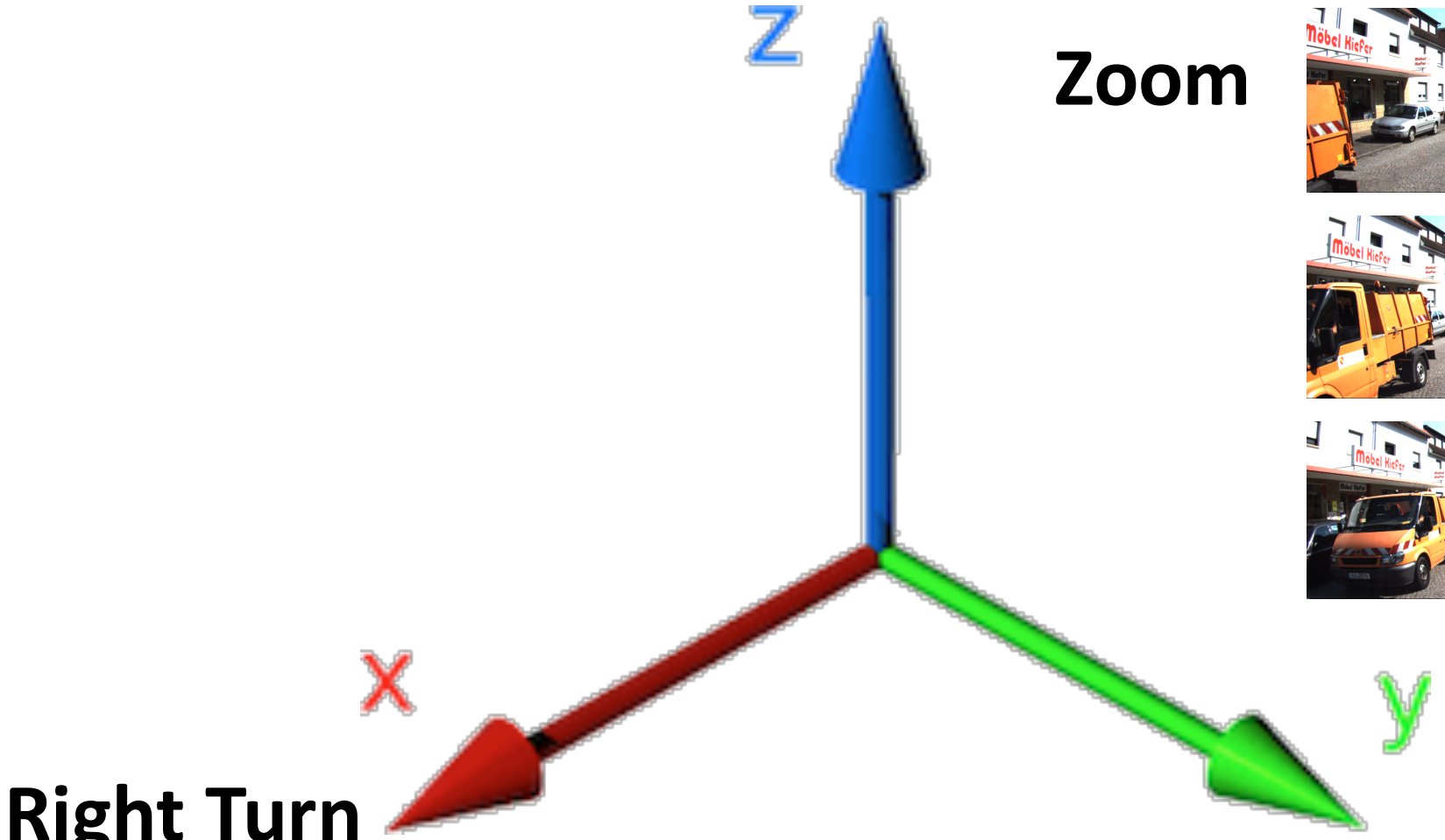
Invariance vs Equivariance



Invariance vs Equivariance



Invariance vs Equivariance



Zoom

Right Turn

Left Turn

Visualization of Learnt Features

Two work about feature learning using ego-motion at ICCV 2015

Learning Image Representations Tied to Ego-motion

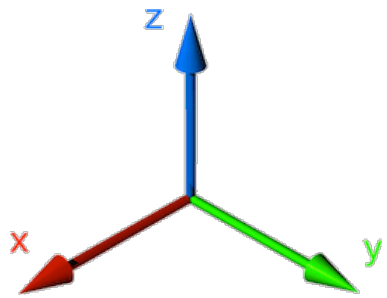
- Dinesh Jayaraman and Kristen Grauman, UT Austin

Learning to See by Moving

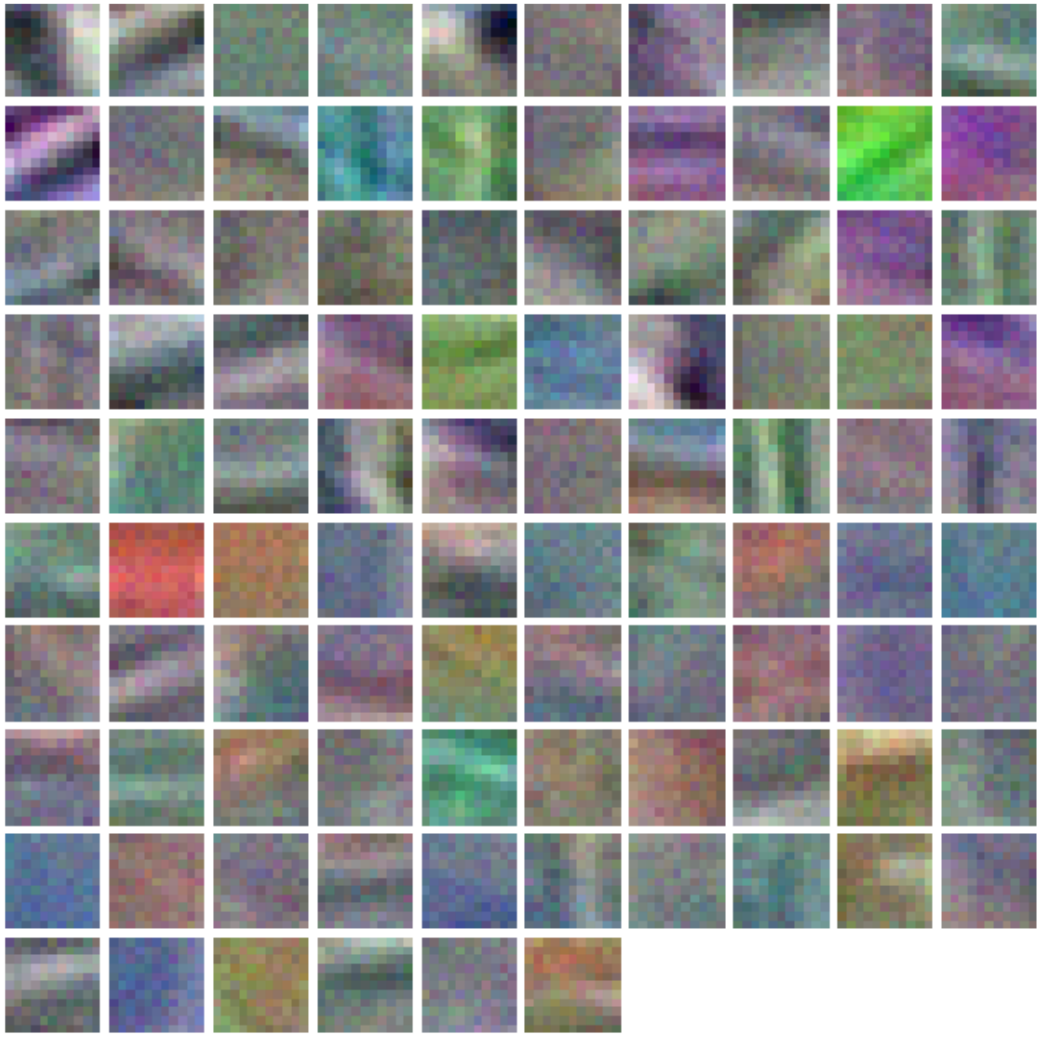
- Pulkit Agrawal, Joao Carreira and Jitendra Malik, UC Berkeley

Visualization of Learnt Features

- KITTI Dataset
  - Learning Image Representations Tied to Ego-motion
    - Ego-motion pattern: K-means clustering based on “yaw” and “forward position”
  - Learning to See by Moving
    - Binning based on three dimensions of camera transformation

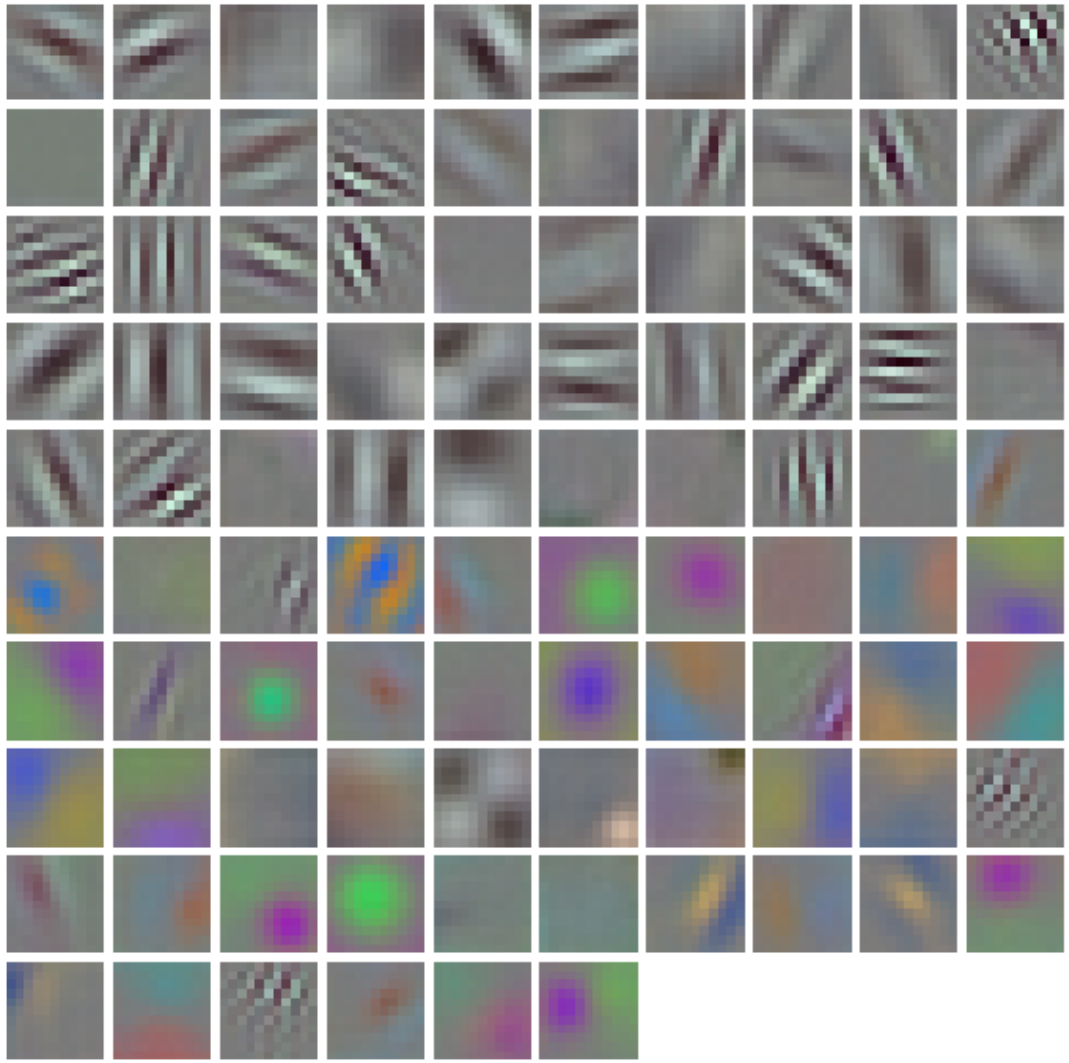


Visualization of Learnt Features



Conv1 Features learnt from KITTI Videos using ego-motion as supervisory signals

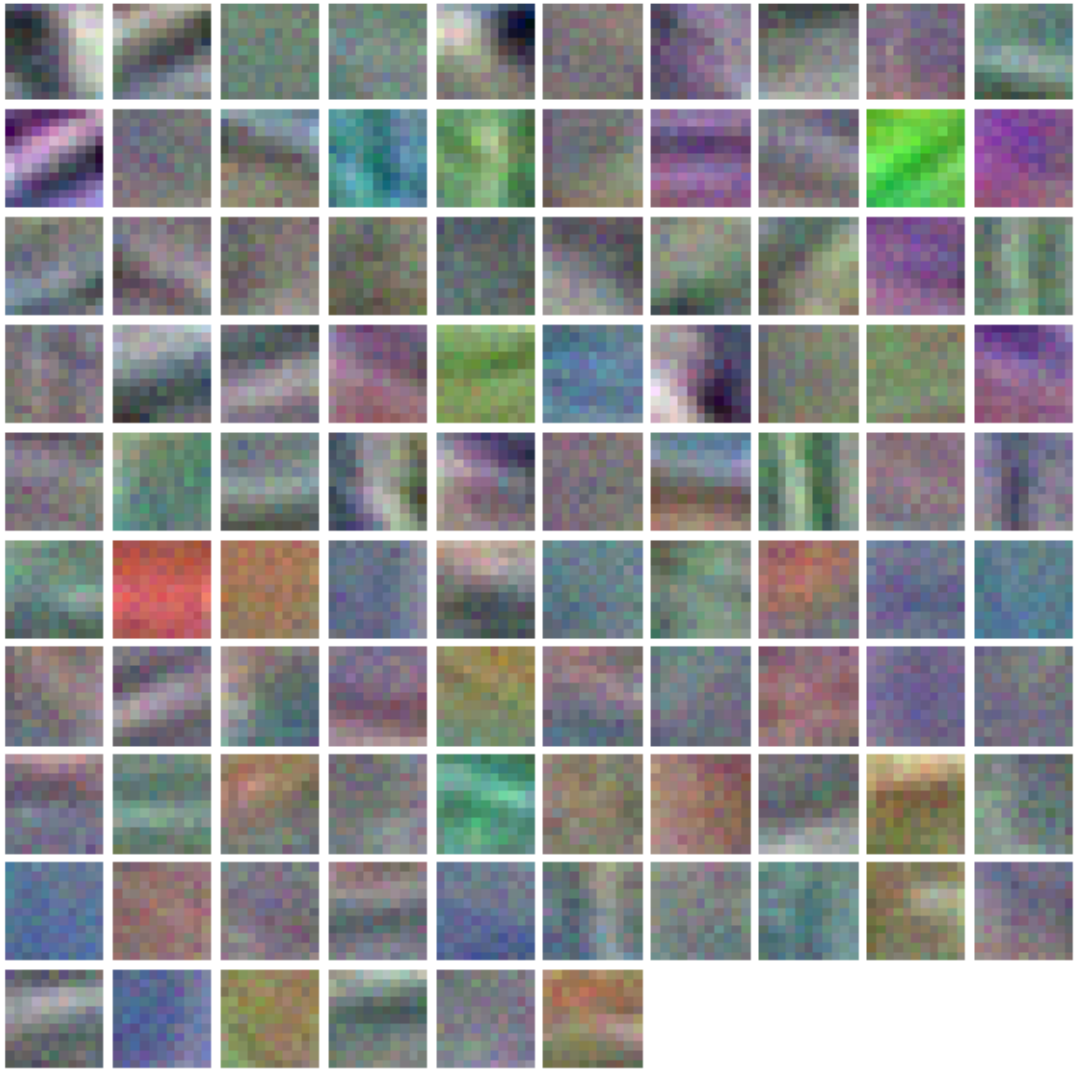
Visualization of Learnt Features

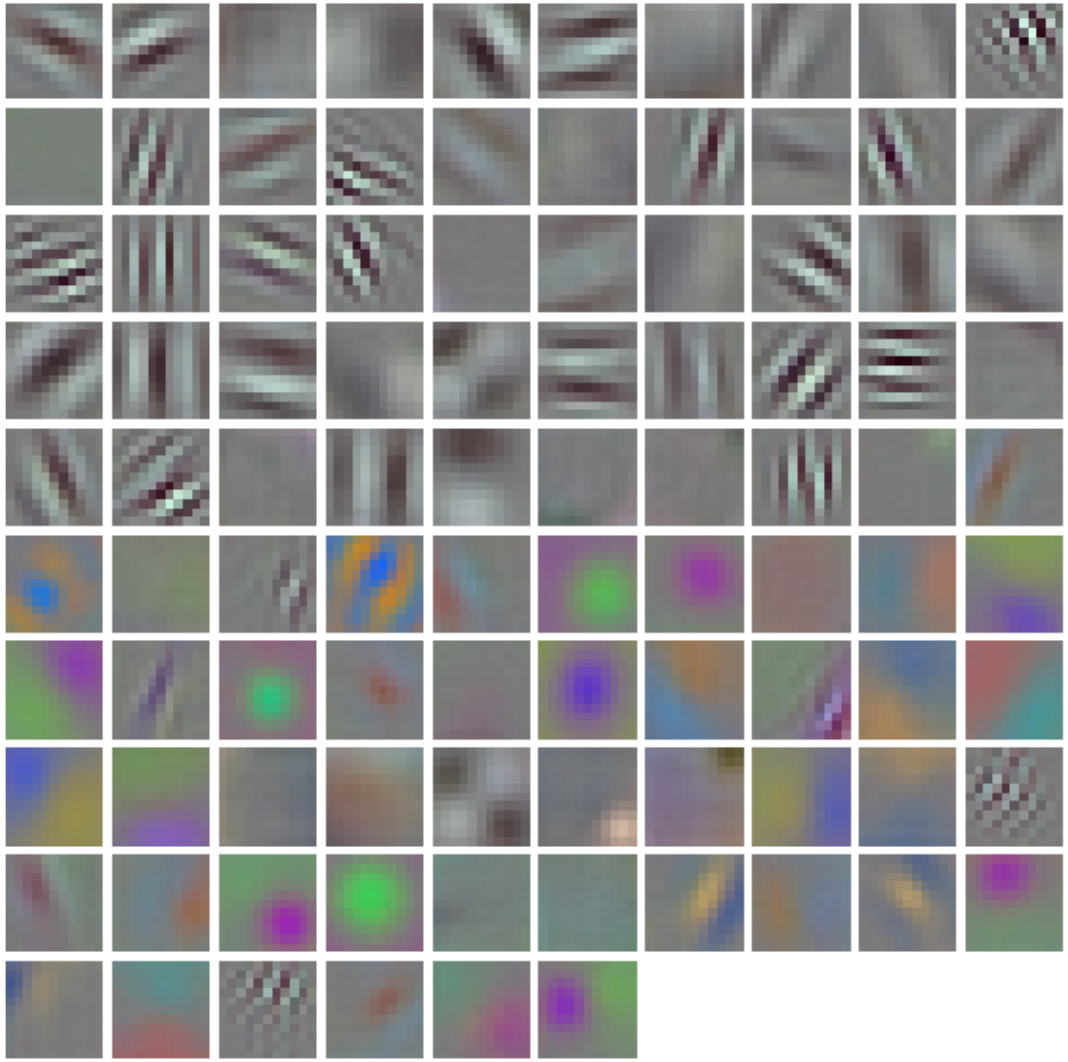


Conv1 Features learnt from labeled images in IMAGENET

Conv1 Features of Pre-trained Alexnet

Visualization of Learnt Features

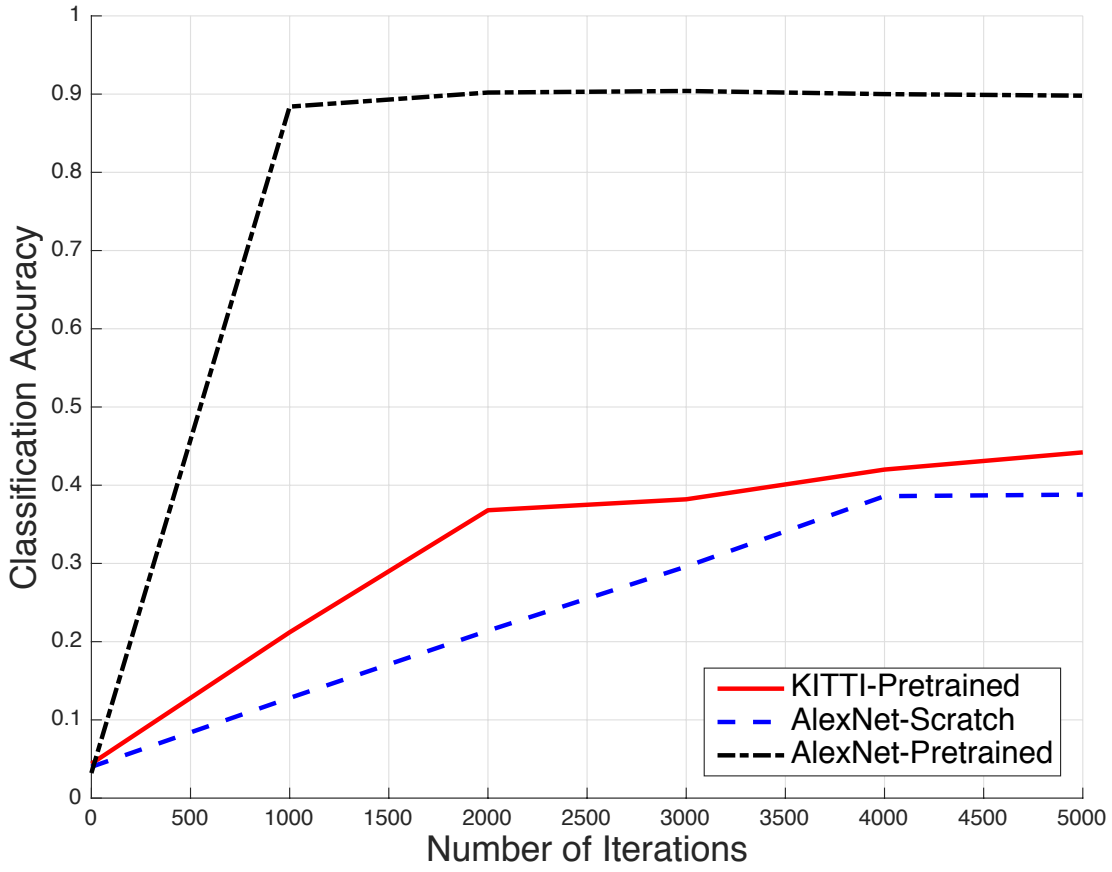




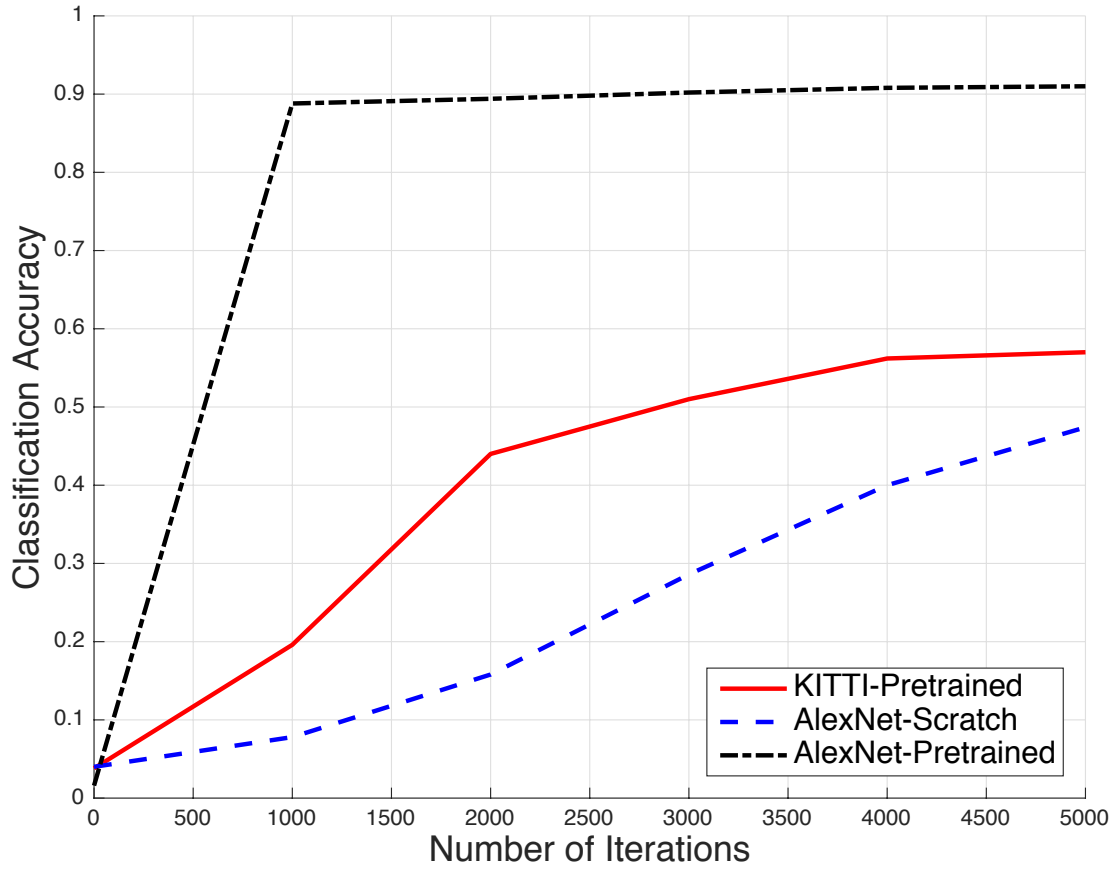
Fine-tuning on New Task

- Task: Assignment 2
- Three Models:
  - AlexNet using randomly initialized weights for all layers
  - AlexNet with the first five layers initialized with features learnt from KITTI Dataset
  - Pre-trained AlexNet model using all ImageNet images
- Fine-tuning setting:
  - Batch size: 50
  - Iterations: 5,000
  - Learning Rate: 0.001
  - momentum: 0.9

Fine-tuning on New Task



2,500 training images



31,500 training images

Reference

1. Dinesh Jayaraman and Kristen Grauman, Learning Image Representations Tied to Ego-motion, ICCV 2015
2. Pulkit Agrawal, Joao Carreira and Jitendra Malik, Learning to See by Moving, ICCV 2015
3. Laurens van der Maaten and Geoffrey Hinton, Visualizing Data using t-SNE
4. Caffe Official Tutorial: http://caffe.berkeleyvision.org/
5. KITTI: http://www.cvlibs.net/datasets/kitti/

Thanks!