Learning handwritten digits with a neural net

MNIST database: 3000 28x28 images of handwritten digits

Sample image:

- Start with random initial weights and use back-propagation to learn weights to recognize digits.
- One output unit for each digit.
- Select output unit with maximum response, e.g., 9.

State-of-the-art recognition systems are based on convolutional neural networks

Public databases of face images serve as benchmarks:
  - >13,000 images of celebrities, 5,749 different identities
  - 3,425 videos, 1,595 different identities

Private face image datasets:
- (Facebook) Social Face Classification dataset
  - 4.4 million face photos, 4,030 different identities
- (Google) 100-200 million face images, ~8 million different identities

<table>
<thead>
<tr>
<th></th>
<th>LFW</th>
<th>YTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook DeepFace</td>
<td>97.4%</td>
<td>91.4%</td>
</tr>
<tr>
<td>Google FaceNet</td>
<td>99.6%</td>
<td>95.1%</td>
</tr>
<tr>
<td>Human performance</td>
<td>97.5%</td>
<td>89.7%</td>
</tr>
</tbody>
</table>

“Deep” neural networks

- Early work extended simple neural networks to have multiple, highly-connected hidden layers.
- If such networks could be trained, they would be much more powerful than “shallow” neural nets.
- But generic multi-layer networks are extremely hard to train!!

Convolutional Neural Networks (CNNs)

Fei-Fei Li, Justin Johnson, Serena Yeung (http://cs231n.stanford.edu/)
Sample stages of a CNN

CONV: "convolution" layer with weights that are learned
RELU: "rectified linear unit" applies an activation function
POOL: "pooling" selects maximum value in small neighborhoods
FC: "fully-connected" neural network

ReLU & max pooling layers

Rectified Linear Activation

Convolutional layer

fully-connected network
locally-connected network

early layers perform a convolution of their inputs
multiple convolution operators (e.g. red & black)
weights in convolution operators are learned
convolution operators are typically small (e.g. 5x5)

Adding a fully-connected neural net layer

Recognizing digits from the MNIST database with a CNN:

LeNet

LeCun, Bottou, Bengio, Haffner (1998)
**AlexNet, ZF Net, GoogLeNet, VGGNet, ResNet, ...**

[Image of AlexNet architecture]

**AlexNet**: Krizhevsky, Sutskever, Hinton (2012)

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Annually since 2010

Maximally activating images from some POOL5 neurons of AlexNet (Girshick et al., 2014)

---

**Facebook’s DeepFace system**

Taigman et al., 2014

- detect face
- 2D align face in crop window using 6 fiducial points
- align to 3D shape model using 67 fiducial points
- use 3D model + image to generate frontal view

---

**Google’s FaceNet system**

Schroff et al., 2015

FaceNet also uses a deep convolutional network

- learns mapping from images to a space where distance between images captures similarity
- training data: triplets of face thumbnails
  - two same ID, one different ID
- learning process: minimize distance between anchor & positive images (same ID), maximize distance between anchor & negative images

Threshold = 1.1 classifies pairs correctly (smaller value means more similar)