Is Bigger CNN Better?

Samer Hijazi on behalf of IPG CTO Group
Embedded Neural Networks Summit (eNNS2016)
San Jose
Feb. 9th
Today’s Story

• Why does CNN matter to the embedded world?
• How to enable CNN in embedded devices?
• Complexity vs. performance tradeoff?
• Concluding thoughts
Why does CNN matter to the embedded world?
What is the semiconductor industry asking for?

We are trying to motivate bridging the gap between academic developments and the semiconductor industry needs.
CNN is growing fast

• Today’s deep learning industry motto is “Bigger is Better”

<table>
<thead>
<tr>
<th>Net name</th>
<th>Layers</th>
<th>Parameters</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet-5 for MNIST (1998)</td>
<td>7</td>
<td>58,996</td>
<td>77,484 M</td>
</tr>
<tr>
<td>ImageNet (2012)</td>
<td>8</td>
<td>60 M</td>
<td>1.1 G</td>
</tr>
<tr>
<td>Deepface (2014)</td>
<td>8</td>
<td>&gt;120 M</td>
<td>1.4 G</td>
</tr>
<tr>
<td>Ensemble CNN (2014)</td>
<td>16x20</td>
<td>23 M</td>
<td>1.4 G</td>
</tr>
<tr>
<td>FaceNet (2014)</td>
<td>22</td>
<td>140 M</td>
<td>1.6 G</td>
</tr>
<tr>
<td>VGG for face (2015)</td>
<td>37</td>
<td>29.7 M</td>
<td>15.5 G</td>
</tr>
</tbody>
</table>

• Embedded devices’ power, price, and form-factor requirements can not accommodate this trend

Dr. Stephen Hicks, Nuffield Department of Clinical Neurosciences, University of Oxford
Complexity vs Performance

- Performance
- Target recognition rate
- Current Industry Trend
- Embedded Device Budget
- Cloud Budget
Methodology

• Step 1. Train a network to your heart’s desire
  – Use the tool/programming language you are most familiar with (e.g., caffe, tensor flow, matcovnet, theano, CNTK, torch, …)

• Step 2. Iteratively reduce the key parameters
  – Utilize:
    – Statistics
    – Linear algebra
  – Use the validation set as guidance for convergence
Balancing performance and complexity
How to do it?

- Reducing number of feature maps
  AlexNet has up to 384 feature maps.  
  \textit{Do we need them all?}

- Reducing network depth
  Oxford VGG face recognition network has 37 layers.  
  \textit{Are they all equally important?}

- Divide and Conquer
  \textit{How many classes should a single network handle?}  
  \textit{What can a single device see?}
Feature Map Reduction
Feature Maps

Number of Feature Maps = Number of 3-D filters
• These filters are highly correlated.
Redundancies in Filter Weights

Q: How many filters do we need?
A: How many independent filters can we really have?

Back to linear algebra 101
Linear Algebra 101

• The space spanned by a vector of size $N$ is referred to as $N$-dimensional space

• Any vector in this space can be expressed as a linear combination of the basis of the space

• It can be shown that there is only $N$ independent basis in the space

\[ \mathbf{a}_i = \sum_{j=1}^{N} c_j \mathbf{v}_j \]
Q: How to find \( v_j \)'s?
A: SVD (Singular Value Decomposition)

\[
A = U \Sigma V^T
\]
Eigen-basis Size Reduction for the Convolutional Filters

- The $i$-th convolutional layer $C^{\downarrow i}$ is of size $L^{\downarrow i} \times W^{\downarrow i} \times H^{\downarrow i} \times K^{\downarrow i}$
  - $L^{\downarrow i}$, $W^{\downarrow i}$, $H^{\downarrow i}$: 3D filter size
  - $K^{\downarrow i}$: number of feature maps
- Form the filter coefficients into a matrix $A$ of size $M \times K^{\downarrow i}$
- Apply SVD decomposition: $A = U \Sigma V^T$
- Use the $M'$ dominating basis (associated with the $M'$ largest singular values) to initialize the filter at $C^{\downarrow i}$
Results Analysis – Traffic Sign Recognition

- German Traffic Sign Recognition Benchmark (GTSRB)
  - 51840 images of German road signs in 43 classes
  - Size of Images varies between 15x15 to 222x193
  - Images grouped by Class and Track with at least 30 images per track.
  - Images available as Color Images (RGB), HOG features, Haar features and Color Histograms.

Figure 3: Random representatives of the 43 traffic sign classes in the GTSRB dataset.
Results Analysis – Traffic Sign Recognition (Cont’d)

- Eliminating the redundancy in the coefficients
  - Improve training efficiency and performance simultaneously
  - Control performance degradation as a function of complexity reduction
Results Analysis – Handwriting Recognition

- The database is provided by Mixed National Institute of Standards and Technology (MNIST)
  - 60,000 training images and 10,000 testing images
  - Black and white images
  - Size normalized to fit in a 20 by 20 pixel box
  - Centered in a 28 by 28 field
• Eliminating the redundancy in the coefficients
  – Improve training efficiency and performance simultaneously
  – Control performance degradation as a function of complexity reduction
Network Depth Reduction
Network Depth

• Q1: Are all layers equally important?

• Q2: How to utilize the difference among layers’ importance to reduce complexity?
Are all layers equally important?

- Define a metric to measure the amount of refinement each layer can contribute.
How to use the layer quality (LQ) metric?

- Eliminate low LQ layers and distribute their functionality over the entire network.
How to use the layer quality (LQ) metric?

• Eliminate low LQ layers and distribute their functionality over the entire network.
  – This can be tricky due to the nonlinear components of the network.

• An alternative way to use LQ metric is to allocate the computational resources based on LQ
Divide and Conquer

Presenter Name and Title (Arial 16pt)
Event Name
Location
Date
What is this?

An animal?

Wolf?
What is this?

An animal?
Wolf?
Cute Wolf?
What is this?

Barking Husky

Howling Wolf
Hierarchical Recognition Concept

- Not all images are equally hard to recognize
- Dynamically allocate resources based on the problem difficulty.
- Let CNN automate the partitioning process.
GTSRB – Ideal Traffic Sign

- Speed Limit Signs: 20, 30, 50, 60, 70, 80, 100, 120
- Other Prohibitory Signs
- Derestricion Signs
- Mandatory Signs
- Other Prohibitory Signs
Hierarchical-CNN for TSR

- Using hierarchical CNN we are able to dynamically allocate resources in accordance with the problem difficulty
- The CNN-oriented family clustering method groups signs into subsets that is more preferable than the pre-defined subset for a CNN classifier.

5+1 different CNNs for classification,
each has lower complexity than one-vs-all classifier
## Results

<table>
<thead>
<tr>
<th>CCR%</th>
<th>Team</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.82</td>
<td>Cadence</td>
<td>HCNN + color</td>
</tr>
<tr>
<td>99.65</td>
<td>Tsinghua</td>
<td>Hinge Loss Trained CNN</td>
</tr>
<tr>
<td>99.46</td>
<td>IDSIA</td>
<td>Committee of CNNs</td>
</tr>
<tr>
<td>99.30</td>
<td>Cadence</td>
<td>Modified UYN</td>
</tr>
<tr>
<td>99.22</td>
<td>INI-RTCV</td>
<td>Human (best individual)</td>
</tr>
<tr>
<td>99.17</td>
<td>Sermanet (UNY)</td>
<td>Updated multi-scale CNN</td>
</tr>
<tr>
<td>98.84</td>
<td>INI-RTCV</td>
<td>Human (average)</td>
</tr>
<tr>
<td>98.31</td>
<td>Sermanet (UNY)</td>
<td>Multi-scale CNN</td>
</tr>
<tr>
<td>96.14</td>
<td>CAOR)</td>
<td>Random Forest</td>
</tr>
<tr>
<td>95.68</td>
<td>INI-RTCV</td>
<td>LDA (HOG 2)</td>
</tr>
</tbody>
</table>
Hierarchical decision process can be incorporated in the network training if we allow gradual decision after each layer.
Closing Thoughts
What is the semiconductor industry asking for?

We are trying to motivate bridging the gap between academic developments and the semiconductor industry needs.
See you in Vegas!

• The premier summer conference for the computer vision community *Computer Vision Pattern Recognition* (CVPR2016) will be held in Vegas from June 27-30.

• Cadence will be offering a tutorial on low complexity recognition SoCs.
  – Embedded CNN
  – Local vs Cloud Compute, why and why not?
  – Complexity Reduction Techniques