Qualcomm Research Deep Net for Video Concept Detection

November 16, 2015
Summary

The Qualcomm Research system is deep learning only
## Inspiration from ImageNet

### Very deep convolutional neural networks

<table>
<thead>
<tr>
<th>Inception</th>
<th>VGGNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small 1x1 convolutions</td>
<td>Small 3x3 convolutions</td>
</tr>
<tr>
<td>Convolution stride of two or one</td>
<td>Convolution stride of one</td>
</tr>
<tr>
<td>ReLU non-linearity</td>
<td>ReLU non-linearity</td>
</tr>
<tr>
<td>Four max-pool layers</td>
<td>Five max-pool layers</td>
</tr>
<tr>
<td>One fully connected layer</td>
<td>Three fully-connected layers</td>
</tr>
<tr>
<td>Dropout</td>
<td>Dropout</td>
</tr>
<tr>
<td>Nine inception modules</td>
<td></td>
</tr>
<tr>
<td>Batch normalization</td>
<td>Simonyan &amp; Zisserman. ICLR 2015</td>
</tr>
</tbody>
</table>

Szegedy et al. CVPR 2015
Batch normalization

- Address covariate shift per layer
- Normalize the activations in each layer within a mini-batch
- Learn the mean and variance of each layer as parameters
- Multi-layer CNN’s train faster with fewer data samples
- Employ faster learning rates and less network regularizations.

- Achieves state-of-the-art on ImageNet, post-competition

\[
\begin{align*}
\mu_B & \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i & & \text{// mini-batch mean} \\
\sigma_B^2 & \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 & & \text{// mini-batch variance} \\
\hat{x}_i & \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} & & \text{// normalize} \\
y_i & \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) & & \text{// scale and shift}
\end{align*}
\]
Approach
High-level overview
All models are pre-trained on ImageNet

- 1,000 standard ImageNet categories
- 1,024 categories better matching the video concepts
- 2,048 same as above, plus 1,024 random categories
- 4,096 same as above, plus more random categories
Data augmentation

Adding color casting and vignetting to default translation and mirroring

Original  Translate/Mirroring  Color casting  Vignetting  All augmentations
Fine-tune

- Inception networks typically have an average pooling on top, making them less suited for domain transfer
  - We add an ‘Alex-style’ fully connected head on the one-but-last layer

- We fine tune the fully connected layers with video labels
  - For both VGGNet and Inception
Video labels

• Common annotation effort finished in 2013  
  Ayache & Quénot, ECIR 2008

• Deep learning profits from more labeled data
  – Relied on Euvision annotations from 2014
  – Hired annotators to correct and supplement
Fusion

- Our models exploit diversity in
  - Networks
  - Image labels
  - Augmentations
  - Video labels

- We have a total of 63 models available for fusion
  - Non-weighted late fusion
  - Weighted late fusion
Experiments
Internal validation set

- Training set
  - 2012devel
  - 2013test
  - 2014test
- Validation set
  - 2012test

<table>
<thead>
<tr>
<th>MediaMill TRECVID 2014 Baselines</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single deep network</td>
<td>56.0</td>
</tr>
<tr>
<td>Seven deep networks</td>
<td>58.0</td>
</tr>
<tr>
<td>Seven deep networks, plus color Fisher vector</td>
<td>60.0</td>
</tr>
</tbody>
</table>
Value of annotations

Additional annotations do not necessarily improve the detection
## Value of image labels

<table>
<thead>
<tr>
<th>Pre-training for single inception model</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 ImageNet baseline</td>
<td>62.2</td>
</tr>
<tr>
<td>1,024 ImageNet for TRECVID</td>
<td>61.7</td>
</tr>
<tr>
<td>2,048 ImageNet for TRECVID + Random</td>
<td>63.1</td>
</tr>
<tr>
<td>4,096 ImageNet for TRECVID + Random</td>
<td>62.3</td>
</tr>
</tbody>
</table>

*Default 1,000 ImageNet categories not necessarily best*
Value of additional data augmentations

<table>
<thead>
<tr>
<th>Model</th>
<th>Default Augmentation</th>
<th>Additional Augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>62.3</td>
<td>63.1</td>
</tr>
<tr>
<td>VGGNet</td>
<td>61.1</td>
<td>61.5</td>
</tr>
</tbody>
</table>

*Additional augmentations give a small but consistent improvement*
## Value of fusion

<table>
<thead>
<tr>
<th>Runs</th>
<th>Fusion</th>
<th>Internal mAP</th>
<th>TRECVID mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gargantua</strong></td>
<td>Non-weighted fusion – all 63 networks</td>
<td>66.9</td>
<td>36.0</td>
</tr>
<tr>
<td><strong>Mann</strong></td>
<td>Weighted fusion – all 63 networks</td>
<td><strong>67.3</strong></td>
<td>35.9</td>
</tr>
<tr>
<td><strong>Edmunds</strong></td>
<td>Non-weighted fusion – 32 networks</td>
<td>66.9</td>
<td>34.9</td>
</tr>
<tr>
<td><strong>Miller</strong></td>
<td>Non-weighted fusion – 7 networks</td>
<td>66.5</td>
<td><strong>36.2</strong></td>
</tr>
</tbody>
</table>

*Seven diverse models fused without weights is good choice*
Great for objects, ok for scenes, poor for actions
10-year progress
Four video data set mixtures

Training

TRECVID 2005

Broadcast news

TRECVID 2007

Documentary video

Testing

Broadcast news

Documentary video

Within domain

Cross domain

Snoek & Smeulders, IEEE Computer 2010
2006-2009: Performance doubled in just three years
2009-2015: same jump by deep learning

Mean average precision
Concept detection on mobile
Qualcomm Zeroth provides on-device deep learning solution
Conclusions

- Deep learning for images leading in video as well
- Technology available on mobile

- TRECVID instrumental in decade of concept detection progress
- Time for a new challenge!
Thank you

Follow us on: f  t  i

For more information on Qualcomm, visit us at:
www.qualcomm.com & www.qualcomm.com/blog