Deep Neural Networks for Automatic Image Annotation

Roman Zakharov
BigData Conference, Kiev Sep'15
Problem definition

When a user takes a photo, the app should check whether they're in a national park...

Sure, easy GIS lookup. Gimme a few hours.

... And check whether the photo is of a bird.

I'll need a research team and five years.

In CS, it can be hard to explain the difference between the easy and the virtually impossible.

PARK?  ???
No idea. There's no GPS info in that photo.

BIRD?  YES
Dude, that is such a bird.

http://parkorbird.flickr.com
Detection vs. Classification

Image detection:
- Detect objects and assign class labels to them

Image classification:
- Assign the correct class label to the whole image

Examples:
- ✓ lamp
- ✓ sofa
- ✓ flower pot
- ✗ painting
Objects in context produce meaning

*FC_name* did it!!

wow!!

hot $)
Image recognition research

Microsoft COCO:
- 328K images
- 2.5M labeled objects
- 91 object type
- 70K worker hours

CIFAR100: 100 classes, 600 images each

PASCAL VOC & ILSVRC challenges datasets

<table>
<thead>
<tr>
<th></th>
<th>PASCAL VOC 2012</th>
<th>ILSVRC 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td># object classes</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td># images</td>
<td>5717</td>
<td>395909</td>
</tr>
<tr>
<td># objects</td>
<td>13609</td>
<td>345854</td>
</tr>
</tbody>
</table>
PASCAL VOC object detection

Before the successful application of ConvNets

After

< 2 years
1.8x mAP

~5 years

Precision: higher is better
Convolutional neural networks

AlexNet, 2012

GoogleLeNet, 2015
GPU-accelerated computing
Deep Learning Magic

convolution + nonlinearity  max pooling

convolution + pooling layers

vec

fully connected layers

Nx binary classification

Low-level: Pool₁

High-level: FC₆
Google did it. Can we?

1. **Caffe**: Deep learning framework by the BVLC
   - convert data
   - define network structure
   - configure solver
   - run and wait

2. **Cloud**: Amazon AWS GPU instance

<table>
<thead>
<tr>
<th></th>
<th>GPU instances</th>
<th>g2.2xlarge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GPU instances</td>
<td>g2.8xlarge</td>
</tr>
<tr>
<td>Memory_optimized</td>
<td>r3.large</td>
<td></td>
</tr>
</tbody>
</table>

3. **Data Science**: educated guess what to do
Problem attack strategy

~2-3K / image of region proposals
(e.g., selective search, EdgeBoxes, MCG)

region-based convolution network

apply trained network to industrial tasks

known datasets of annotated single & multi-class images:
ImageNet, VOC, CIFAR, COCO...
It takes some time and fun
Results I
Results II
Thank you!

any questions?

roman.zakharov@softcube.com