Vision, Feedback and Egomotion

João Carreira
Google Deepmind

University of Oxford
22-02-2016
Change of plans relative to advertised abstract

- Learning to See by Moving Agrawal, me and Malik. ICCV 2015.

Pulkit Agrawal  Katerina Fragkiadaki  Jitendra Malik
Part 1

Contextual Prediction with Iterative Error Feedback
Classification

We can now do it well, using ConvNets!

Dog
ConvNets as Hierarchical, Feedforward Feature Extractors

Image adapted from Martin Riedmiller
In Human Vision Basic Recognition is a Hierarchical, Feedforward Process.

Slide adapted from Jack Gallant
Computer Vision

ConvNets brought large improvements to object localization, semantic segmentation and human pose estimation.

Girshick et al CVPR 2014
Zheng et al 2015
Pfister et al 2015

CRF-RNN
Open Challenges

• Complex output spaces (large scenes, 3D shapes)
Open Challenges

- Complex output spaces (large scenes, 3D shapes)
- Handling occlusions / missing data
Open Challenges

• Complex output spaces (large scenes, 3D shapes)
• Handling occlusions / missing data
• Continuous processing in time
What’s missing

• Complex output spaces (large scenes, 3D shapes)
• Handling occlusions / missing data
• Continuous processing in time

Keeping context
Keeping Context - Iterative Feedback

1. Feed back predictions
2. Render them as an “image”
3. Concatenate with real image
4. Feed forward through ConvNet
Keeping Context - Iterative Feedback
Keeping Context - Iterative Feedback
Keeping Context - **Iterative Error Feedback**

Error of Previous State

Classes

Object Parts

Contours

Edges

Pixels

Previous State

New State

15
Context in the Brain: Feedback Connections
Context in the Brain: Feedback Connections

- Lateral Geniculate Nucleus usually assumed to be a simple relay to visual cortex
- But **80-95%** of Lateral Geniculate Nucleus **feedback** from Visual Cortex
- It is unknown why.
Running Example: Human Pose Estimation

- Determine 2D positions of set of keypoints
Running Example: Human Pose Estimation

1. Render pose as set of image channels

- Right Foot
- Head
Running Example: Human Pose Estimation

1. Render pose as set of image channels
Running Example: Human Pose Estimation

1. Render pose as set of image channels
2. Stack pose channels with image channels to form ConvNet inputs
Running Example: Human Pose Estimation

1. Render pose as set of image channels
2. Stack pose channels with image channels to form ConvNet inputs
3. Predict bounded correction for current pose
4. Update pose and goto 1
Iterative Error Feedback for Human Pose

Only 3 keypoint channels shown

Initialize with mean pose
Human Pose Estimation – Example Image
Human Pose Estimation – Example Image
Human Pose Estimation – Example Image
Human Pose Estimation – Example Image
Human Pose Estimation – Example Image
Learning

- Regular feedforward training. Each iteration is a separate training example.
- Linear paths between key point in initial and ground truth pose.
- Bounded maximum displacements.
- Initial pose = ground truth pose from random different image.
Learning

• Example training examples for one image
Learning

• Example training examples for one image
Human Pose Estimation – MPII Dataset

• MPII Human Pose Dataset
  • 25k images and 40k annotated people.
  • One point in each person’s torso provided.

Hardest and most realistic existing dataset for 2D human pose estimation
Results given ground truth person size

PCK-h at 0.5 overlap
Larger is better (maximum 100.0)

<table>
<thead>
<tr>
<th>Method</th>
<th>Full Body Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang and Ramanan</td>
<td>44.5</td>
</tr>
<tr>
<td>Tompson et al</td>
<td>82.0</td>
</tr>
<tr>
<td>Regression-googlenet</td>
<td>74.8</td>
</tr>
<tr>
<td>IEF-googlenet</td>
<td>83.4</td>
</tr>
</tbody>
</table>

Articulated human detection with flexible mixture of parts. Yang and Ramanan, TPAMI 2013
Efficient Object Localization using Convolutional Networks. Tompson, Goroshin, Jain, Lecun and Bregler, CVPR 2015
Dealing with scale

• Train a ConvNet to pick tightest window, using ground truth annotations.
Results without ground truth scale

PCK-h at 0.5 overlap
Larger is better (maximum 100.0)

<table>
<thead>
<tr>
<th>Method</th>
<th>Full Body Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tompson et al on original image scale</td>
<td>66.0</td>
</tr>
<tr>
<td>IEF on original image scale</td>
<td>76.0</td>
</tr>
<tr>
<td>IEF with scale predictor</td>
<td>83.4</td>
</tr>
<tr>
<td></td>
<td>Regression all Joints</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Left Knee PCKh-0.5</td>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Wrist PCKh-0.5</td>
<td>61.8</td>
</tr>
</tbody>
</table>
Learning $g()$

- Preliminary experiment replacing the gaussians by an upsampling ConvNet.
- Just 1% worse compared to model with gaussians, on first attempt.
- Important for generalizing to a more general set of tasks.
Human Pose Estimation – Results

Initialization  Intermed. Prediction  Final Prediction  Ground Truth
Human Pose Estimation – Results

Initialization  Intermed. Prediction  Final Prediction  Ground Truth
Human Pose Estimation – Results

Initialization  Intermed. Prediction  Final Prediction  Ground Truth
Human Pose Estimation – Results

Initialization  Intermed. Prediction  Final Prediction  Ground Truth
The Down Side

- Less competitive PCK-h at more stringent thresholds.
- Recent (2 weeks ago) method “Convolutional Pose Machines” from CMU gets 87.8 vs our current 83.4.
The Down Side

• Less competitive PCK-h at more stringent thresholds.
• Recent (2 weeks ago) method “Convolutional Pose Machines” from CMU gets 87.8 vs our current 83.4.

but ResNet is now available for MatConvNet (training …)!
Part 2
Learning to See by Moving
*One* Feedforward Feature Extractor

• 1970s
  • Many integrated circuits for many different tasks. Then rise of the microchip. One integrated circuit for many different tasks.

• Same may happen now for visual system architectures
  • What is the best architecture? What task to train it on?
What is a good generic feature extractor?

- Should allow us to:
  1. Perform multiple tasks

- Recognize places
- Recognize objects
- Establish correspondences
- Visual Odometry
What are Useful Features?

- Should allow us to:
  1. Perform multiple tasks
  2. Learn to perform new tasks with minimal “external” supervision
ImageNet Classification ConvNets

- Trained with ~1M web images on the task of image classification after manually annotating class labels
Visual Development in Biology

• Babies are not shown 1M web images with annotated labels. How do they learn?
Visual Development in Biology

• Babies are not shown 1M web images with annotated labels. How do they learn?
Perception – Action Loop

Traditional view: How can perception inform actions
Perception – Action Loop

Traditional view: How can perception inform actions

Our work: how can actions inform perception
Egomotion

• Simplest action: moving around in an environment
• Mobile agents have non-visual access to egomotion
Egomotion

- Simplest action: moving around in an environment
- Mobile agents have non-visual access to egomotion
- **Learning principle**: visual features should allow predicting egomotion caused by agent’s motor system
Learning Architecture

Base-CNN Stream-1

\[ L_1 \rightarrow L_2 \rightarrow \ldots \rightarrow L_k \]

Base-CNN Stream-2

\[ L_1 \rightarrow L_2 \rightarrow \ldots \rightarrow L_k \]

Top-CNN

\[ F_1 \rightarrow F_2 \]

Transformation
Testing Architecture
Evaluating Utility of Features

• Pretrain with
  • **Egomotion (this work)**
  • Class-based supervision (Krizhevsky et al NIPS 2012)
  • Slow Feature Analysis (Mohabi et al ICML 2009)

• Finetune and evaluate on
  • Scene Recognition
  • Image Classification
  • Keypoint Matching
  • Visual Odometry
Experimental Setup

• KITTI Dataset
  • A car moving in urban scenes
  • Odometry data available
  • ~20K unique frames

• ConvNet Architecture
  • AlexNet
### Scene Classification – SUN397

#### Accuracy - Higher Better

<table>
<thead>
<tr>
<th></th>
<th>5 Finetuning Examples / Class</th>
<th>20 Finetuning Examples / Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex-1M</td>
<td>18.0</td>
<td>33.3</td>
</tr>
<tr>
<td>Alex-20K</td>
<td>6.6</td>
<td>12.5</td>
</tr>
<tr>
<td>Kitti-20K</td>
<td>6.4</td>
<td>12.4</td>
</tr>
</tbody>
</table>

- Imagenet Pretraining
- Imagenet Pretraining
- Egomotion-based Pretraining
Visual Odometry – SF Dataset

SF Dataset

Accuracy - Higher Better

<table>
<thead>
<tr>
<th>Method</th>
<th>Translation Acc.</th>
<th>Rotation Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\delta X$</td>
<td>$\delta Y$</td>
</tr>
<tr>
<td>KITTI-Net</td>
<td>43.4</td>
<td>57.9</td>
</tr>
<tr>
<td>AlexNet-1M</td>
<td>41.8</td>
<td>58.0</td>
</tr>
</tbody>
</table>

Egomotion-based Pretraining
Imagenet Pretraining
Keypoint Matching - PASCAL VOC

Lower Curves → Better Performance
Image Classification Accuracy - Imagenet

Krizhevsky et al NIPS 2012

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet-Scratch</td>
<td>1.1</td>
<td>3.1</td>
<td>5.9</td>
<td>14.1</td>
</tr>
<tr>
<td>KITTI-SFA-Net (Slowness)</td>
<td>1.5</td>
<td>3.9</td>
<td>6.1</td>
<td>14.9</td>
</tr>
<tr>
<td>KITTI-Net (Egomotion)</td>
<td>2.3</td>
<td>5.1</td>
<td>8.6</td>
<td>15.8</td>
</tr>
</tbody>
</table>
Conclusions

• Iterative Error Feedback as a simple approach for adding contextual prediction capabilities to feedforward feature extractors.
  • Rephrasing the prediction problem from “what is there?” to “what changed?”
  • Competitive results on human pose estimation.

• Self-supervised learning of ConvNets by leveraging non-visual sensors.
  • Valid for mobile agents.
  • Promising results but so far explored only small-scale datasets.
Thank you!