CAP 6412 Advanced Computer Vision


Boqing Gong

Feb 18, 2016
Today

• Administrivia
• Neural networks & Backpropagation (VIII)
• Pose estimation, by Amar
Next week: Vision and language

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Suhas Nithyanandappa</td>
<td></td>
</tr>
<tr>
<td>Nandakishore Puttashamachar</td>
<td></td>
</tr>
</tbody>
</table>
Project 1: Due in 10 days (02/28)

• If you have discussed option 2 with me
  • Send me the meeting minutes / slides --- grading criteria

• If you take option 1
  • In total, >6,000 validation images
  • Test 3 images per class of the validation set
Today

• Administrivia

• Neural networks & Backpropagation (VIII)

• Pose estimation, by Amar
Recap

Data: \((x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, 2, \cdots, n\)

Goal: Find the labeling function \(c : \mathcal{X} \rightarrow \mathcal{Y}, c(x) = y\)

Hypotheses: \(\text{NET}(x; \Theta)\)

Expected risk: \(R(\Theta) = \mathbb{E}_{(x,y)}[\text{NET}(x; \Theta) \neq y]\)

Empirical risk: \(\hat{R}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i; \Theta)\)
Recap 2

Empirical risk: \( \hat{R}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i; \Theta) \)

Parameter estimation: \( \hat{\Theta} \leftarrow \arg \min_{\Theta} \hat{R}(\Theta) \)

Optimization by stochastic gradient descent (SGD): \[ w^t \leftarrow w^{t-1} - \eta \nabla L(x_i, y_i; \Theta^{t-1}) \]

\( \eta \): learning rate
Checking for convergence 1

• What happened?
• 1. Bug in the program
• 2. **Learning rate too high**
• 3. What if $c(x)=y$ is totally random?
• 4. No appropriate pre-processing of the Input data $\rightarrow$ dummy gradients
Checking for convergence 1

- What happened?
Checking for convergence 2

- What happened?
  - Different learning rate $\rightarrow$ different local optimum & decreasing rate

- Why oscillation:
  - 1. Skipped the optimum because of large learning rate
  - 2. Gradient of a single data point is noisy
  - 3. Not computing the real $R$, instead we approximate it (see P12)
Checking for convergence 3

• Turning down the learning rate!
• After some iterations:

\[ \eta = \frac{\text{const 1}}{t + \text{const 2}} \]
Computing $\hat{R}$

Empirical risk: $\hat{R}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i; \Theta)$

- Impossible when $n$ is large (e.g., $n=10,000,000$)
- Every $T$ iterations,
  
  average $L(x_i, y_i; \Theta^t)$ over the $T$ iterations
Training, validation, and testing

Training data: \((x_i, y_i), i = 1, 2, \ldots, n\)
Validation data: \((x_i, y_i), i = n + 1, \ldots, n + m\)
Test data: \((x_i, y_i), i = n + m + 1, \ldots, n + m + k\)
Overfitting

- Training, validation, test
- What shall we do?
  - Early stopping
  - Data augmentation
  - Regularization (Dropout)
  - Reduce the network complexity
- Which place should we (early) stop?
  - Theta_1
Under-fitting

- Training, validation, test
- What shall we do?
Today

• Administrivia
• Neural networks & Backpropagation (VIII)
• Pose estimation, by Amar
Upload slides before or after class

• See “Paper Presentation” on UCF webcourse

• Sharing your slides
  • Refer to the original sources of images, figures, etc. in your slides
  • Convert them to a PDF file
  • Upload the PDF file to “Paper Presentation” after your presentation
Flowing ConvNets for Human Pose Estimation in Videos

Tomas Pfister, James Charles, Andrew Zisserman

Presenter: Amar Nair
amarknair@knights.ucf.edu
CONTENTS

• Introduction
• Problem Statement
• Related Work
• Approach Outline
• Details of the Proposed Approach
• Convolutional Architecture
• Experiments
• Results
• Conclusion
• Future Directions
INTRODUCTION

• Still Image Pose Estimation
• Pose Estimation from Videos – Frames provide additional cue for recognition
Applications

- Action Recognition
- Human Computer Interaction
- Surveillance
RELATED WORK

• Local image evidence and relative positions in the kinematic Chain
• Poselets – Detectors trained using 3D pose annotations
• Random Forest on depth data and RGB
• Regress Heatmaps directly for joints
Approach Outline

• Deep Expert Pooling Architecture for Pose Estimation
Spatial ConvNet

- Deep Expert Pooling Architecture for Pose Estimation
Spatial ConvNet

- Regressing heatmap of each joints separately
- L2 Regularisation
- Given training data $N = \{X, y\}$ and the ConvNet regressor
- $\phi$ (output from conv8)

\[
\arg\min_{\lambda} \sum_{(X,y)\in N} \sum_{i,j,k} \| G_{i,j,k}(y_k) - \phi_{i,j,k}(X, \lambda) \|^2 \\
\text{where } G_{i,j,k}(y_i) = \frac{1}{2\pi\sigma^2} e^{-\left[(y_k^1-i)^2 + (y_k^2-j)^2\right]/2\sigma^2}
\]
Spatial Fusion Layers

- The fusion layers learn to encode dependencies between human body parts locations, learning an implicit spatial model.
OPTICAL FLOW FOR POSE ESTIMATION

1. Warping confidence maps with optical flow
2. Pooling the confidence maps
3. Maximum confidence from the composite map
IMPLEMENTATION

• The input frames are rescaled to height 256. A $248 \times 248$ sub-image (of the $N \times 256$ input image) is randomly cropped, randomly horizontally flipped, randomly rotated between $-40^\circ$ and $-40^\circ$ and resized to $256 \times 256$.

• Training Time - Training SpatialNet on FLIC took 3 days & SpatialNet Fusion 6 days.

• Optical Flow - Optical flow is computed using FastDeepFlow with Middlebury parameters.
DATASETS

• BBC Pose Dataset - consists of 20 videos (each 0.5h–1.5h in length) recorded from the BBC.

• Extended BBC Pose Dataset - This dataset contains 72 additional training videos which, combined with the original BBC TV dataset.

• ChaLearn Dataset - Multi-modal gesture dataset contains 23 hours of Kinect data of 27 people.

• Poses in the Wild and FLIC datasets. The Poses in the Wild dataset contains 30 sequences (total 830 frames) extracted from Hollywood movies.
Experiments

1. Evaluation Protocol and Details
2. Component Evaluation
3. Comparison to State of the art

A demo video is available at http://youtu.be/pj2N5DqBOgQ.
Evaluation Protocol and Details

- Evaluation Protocol - estimated joints against frames with manual ground truth (except ChaLearn, where comparison against output from Kinect).
- Experimental Details – All frames are shuffled and augmented to present maximally varying input.
- Baseline Method – CordinateNet trained for regressing the joint positions.
- Computation Time – Real-time, 50fps on 1 GPU without optical flow, 5fps with optical flow.
Component Evaluation

• Comparison of the performance of ConvNets for wrists on BBC Pose
Component Evaluation

• Comparison of pooling types
Comparison to the state of the art

- BBC Pose
Comparison to the state of the art

- ChaLearn
Comparison to the state of the art

- FLIC
Related Work

- Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies
  - Seguin, Guillaume, Karteek Alahari, Josef Sivic, and Ivan Laptev.
Conclusion

• Simple, direct and improved (by combining with optical flow) ConvNet
• Outperforms the state of the art methods on the standard datasets.
FUTURE DIRECTIONS

- Action Recognition in videos
- Object tracking in surveillance
- Train on more datasets