Context-Aware CNNs for Person Head Detection
Tuan-Hung Vu, Anton Osokin, Ivan Laptev
INRIA/ENS, Paris, France

Goal

Context-aware CNNs

Local model

CNN-based detector, trained on Selective Search object proposals [similar to R-CNN [1]]
Pre-training on ImageNet [2]

Fine-tuning the sum of independent log-losses using SGD.

Global model

Predicts positions and scales of objects given the whole image as input
The target is defined over a coarse grid of image regions (cells)
Label each cell as positive if its region has sufficient overlap with a ground-truth bounding box

Training: minimizing the sum of independent log-losses using SGD.

Pairwise model

Similar to [3], we construct the candidates in a given image:

\[ S(y; w) = \sum_{i \in V} \theta^{L}_{i}(y_{i}) + \sum_{(i,j) \in \mathcal{E}} \theta^{P}_{ij}(y_{i}, y_{j}, k_{ij}) \]

where \( V \) is the set of all examined candidates, and \( y = (y_{i})_{i \in V} \)
are the corresponding label assignments, \( \theta \)—trainable parameters

For each candidate \( i \), \( a \) is computed as the difference of the max-marginals of the joint-score:

\[ s_{i}(w) = \max_{a} S(y_{i}; w) - \max_{a} S(y_{i}; w) \]

Logistic loss on the structured scores:

\[ l(w, y, a) = \sum_{d \in D_{i}} v(s_{d}(w)) + \sum_{d \in D} v(-s_{d}(w)) \]

Training step:
1. Construct a set of candidates using local model.
2. Perform forward pass to compute potentials.
3. Perform inference to compute structured loss and its gradient.

Local region-based models do not capture the context

HollywoodHeads dataset

Collected from 21 Hollywood movies of different genres from different time periods
In total: 2,380 clips with 3,872 human tracks spanning over 224,740 frames
Bounding-box annotation for heads on key frames
Linear interpolation and manual verification on all frames
Training: 216,719 frames from 15 movies; validation: 6,719 frames from 3 movies; test: 1,302 frames from 3 movies

Contributions

Propose two Global and Pairwise models
with 369,846 head bounding-box annotations in 224,740 movie frames

Motivation

For person detection, face detectors are insufficient and full/upper body detectors often fail in close-up views
Success of Convolutional Neural Net in object detection
Image context embeds constraints on the global and relative positions of objects in the image
Local region-based models do not capture the context

We validate the method on the new dataset, dataset [4] and dataset [5]. For each dataset we evaluate:

- Local, Local+Global, Local+Pairwise, Local+Pairwise+Global models

Related work