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

Goal
- Improve region-proposal-based CNN [1] with contextual CNNs for human head detection

Local model
- CNN-based detector, trained on Selective Search object proposals [3] and R-CNN [1]
- Pre-training on ImageNet [2]
- Fine-tuning on HollywoodHeads dataset, minimizing the sum of independent log-losses using SGD.
- Additional training on ImageNet [2]

Global model
- Predicts positions and scales of objects given the whole image as input
- The target is defined over a coarse multi-scale 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 C log-loss functions

Context-aware CNNs

Pairwise model
- Similar to [3], we construct the joint score function for object candidates in a given image:

Here V is the set of all examined candidates, and are the corresponding label assignments, U–trainable parameters
- For each candidate i, a score is computed as the difference of the max-marginals of the joint-score
- Structured surrogate loss – logistic loss on the structured scores

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
4. Back-propagate the gradient

Results
- We validate the method on the new HollywoodHeads dataset, TVHI dataset [4] and Casablanca dataset [5]. For each dataset we evaluate:
  - Local, Local+Global, Local+Pairwise, Local+Pairwise+Global models
  - DPM Face detector [6]

Related work

Base architectures:
- Pairwise model: pairwise potentials
- Pair cluster
- Unary potential
- Multi-scale grids
- CNN: feature extractor
- Multi-scale grids
- Local model
- Output of the Global model
- Global + Local model
- Complexit reduction: performance with different candidate-left ratio after filtering using Global Model