Big data
Phuong Pham
April 9, 15
Learning Everything about Anything: Webly-Supervised Visual Concept Learning

Based on Santosh Divvala’s slide
Motivation

What are these?

Rearing horse

Rolling horse

Reining horse

Eye horse
A system that can

Without human supervision

• Biased, non-comprehensive
• Concept-specific expertise
• Scalability
Approach

Overall questions? Or we will go into detail of each component

[Diagram of the approach process]
Retrieve concept’s variations

• Approximately 5000 n-grams per concept
• Several visually non-salient n-grams e.g. “last horse”, “particular horse”
• Classifier-based pruning (5000 -> 1000)
  • Train/test thumbnail images, if A.P. < threshold, discard n-gram
• Find good quality n-gram
Good quality n-grams

\[
\max_S \sum_{i \in V} d_i \cdot \mathcal{O}(i, S)
\]

\[
\mathcal{O}(i, S) = \begin{cases} 
\frac{1}{1 - \prod_{j \in S} (1 - e_{i,j})} & i \in S \\
1 & i \notin S 
\end{cases}
\]

such that \( |S| \leq k \)

Merge into \( S \) if \( i, j \) are different?

Not for finding super-n-grams!
Super-ngrams

Merging similar “good quality” ngrams
Training models

• Train separate DPM per super-ngram
• Pruning noisy components
• Merging similar appearance clustering (~ 50%)
Pruning noisy components

Noisy components (pruned based on A.P. threshold and frequency)
Merging similar components (as super-ngram)
Object detection

- Their system is comparable to WS_vid_CVPR12 (better in avg)
- WS Img ICCV11 is the best model at the cost of fully supervised learning (limitation of annotation)
Action detection

VOC Challenge

Our Goal

Given the person bounding box in an image, identify the action

Given a image, identify and localize the action in an “unsupervised” approach

The webly learning give reasonable good result?
Future work
Summary

• Advantages
  • Cut annotation cost
  • Combine linguistic and visual semantic may lead to more interesting research topics
  • ?

• Disadvantages
  • Still not defeat state of the art even with training set at large scale
  • ?
Questions
Scene Completion using Millions of Photographs

Based on James Hays’ slides
Motivation
The algorithm

Input image → Scene Descriptor → Image Collection

20 completions → Context matching + blending → 200 matches
Semantic scene matching

Gist scene descriptor from 6 orientation and 5 scales
Find its L2 nearest neighbors in 2.3 million images}

→ ... 200 total
Context matching

Graph cut + Poisson blending
Graph cut seam finding

\[ C(L) = \sum_{p} C_d(p, L(p)) + \sum_{p,q} C_i(p, q, L(p), L(q)) \]

For missing regions of the existing image
\[ C_d(p, exist) \]

For regions of the image not covered be the scene match
\[ C_d(p, patch) \]

For other pixels
\[ C_d(p, patch) = (k \times \text{Dist}(p, hole))^3 \]
Result ranking

We assign each of the 200 results a score which is the sum of:

- The scene matching distance
- The context matching distance (color + texture)
- The graph cut cost
Qualitative
Qualitative - failures
Quantitative

• Our: 37%
• Baseline: 10%
Human evaluation
Discussions