Deep Unsupervised Similarity Learning using Partially Ordered Set

Miguel Bautista, Artsiom Sanakoyeu, Björn Ommer

Jens Beyermann

June 7, 2018

Overview

Motivation

Histogram of Oriented Gradients

The Concept

Method

Experiments

Motivation

Motivation

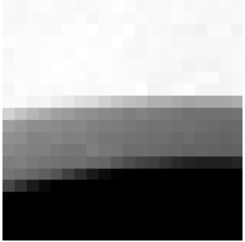
- There is too much data to annotate everything for supervised learning.
- Even if we could label all the data, this would be very costly.
- Unsupervised learning helps to make use of the available data.

Histogram of Oriented Gradients

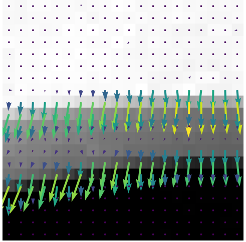
Histogram of Oriented Gradients (HOG)

- HOGs give a feature representation of images.
- They can be visualized intuitively.
- They are a good starting point for our later method since their behaviour is closely connected to the behaviour of CNNs.

Detecting Edges by Image Gradients

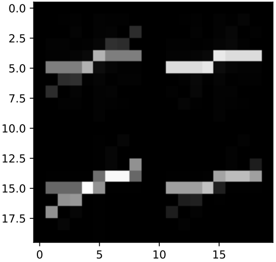


(a) input image

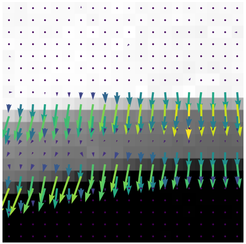


(b) pixelwise gradients

Building Histograms from Image Gradients



(c) histogram of gradients

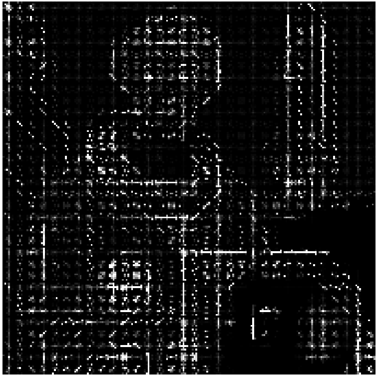


(d) pixelwise gradients

Obtaining the Full HOG

- Repeat this process on a sliding window over the whole image.
- Concatenate all obtained histograms to a long feature vector.
- Linear discriminant analysis can be used to suppress background gradients.

Visualization of the total HOG



Distances Between HOGs

- HOGs are basically d dimensional vectors that represent an image in an abstract feature space.
- We denote The projection from an image x to its HOG-vector $HOG(x)$ with ϕ.
- We can compute the euclidean distance $||\phi(x) - \phi(y)||$ between HOGs of images x and y.
- It turns out that this distances are quite reliable for very close or very distant images, but not for images with a "mediocre" distance.

The Concept

The Concept

- We can represent pictures by their HOGs.
- This allows detection of very similar/unsimilar pictures.
- HOG similarities are unreliable on a “mediocre” scale.

The Concept

- We can represent pictures by their HOGs.
- This allows detection of very similar/unsimilar pictures.
- HOG similarities are unreliable on a "mediocre" scale.

Idea: Use similarity learning with HOG-similarity as starting point, to enhance the performance for images with unclear similarity (mediocre distance).

Surrogate Classification

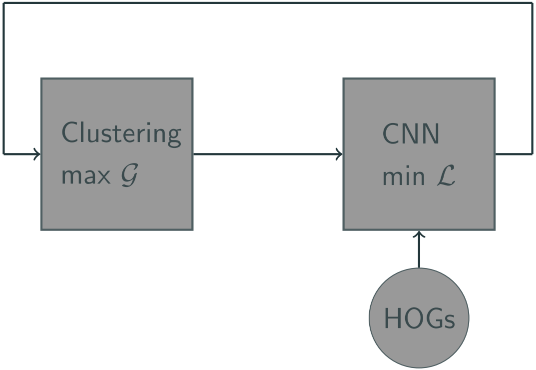
In this approach we implement similarity learning as a "surrogate classification" task.

- We obtain surrogate classes from the reliable (i.e. very close) similarities with clustering.
- We use a CNN to learn a projection into an abstract feature space, that reproduces those classifications.

How to Use Our Data?

Since only few samples in a common dataset are close enough to form a surrogate class, we do not use the vast majority of our data.
Idea: We can use this data if we obtain partial orderings to model more “fine grained” similarities.

Architecture Scheme



Summary

So our method is based on two different steps.

- Learn representations in an abstract feature space, starting with the HOG.
- Compute groupings into surrogate classes based on the distances in the current feature space.

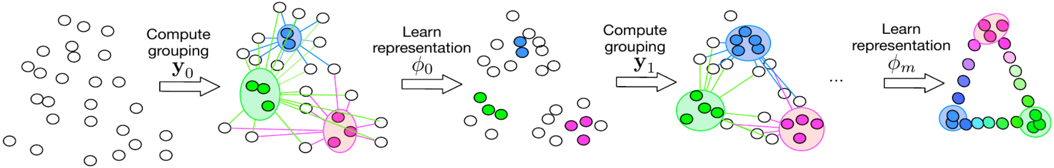
These steps get repeated, by a "joint optimization process" implemented with a convolutional neuronal network with a "recurrent" training process.

Method

Some Notation

- $X = (x_1, \dots, x_n)^T, \; x_i \in \mathbb{R}^p$ is our dataset of images.
- θ are the parameters defining the state of a CNN.
- $\phi^\theta : X \to \mathbb{R}^{1 \times d}$ is the projection into the feature space, represented by the CNN given by θ.

Schematic Method



Partial Ordering (repetition)

Partial Ordering

A partial order is a binary relation $\leq$ over a set X meeting the following requirements $(x, y, z \in X)$:

1. $x \leq x$ (reflexivity)
2. if $x \leq y$ and $y \leq x$, then $x = y$ (antisymmetry)
3. if $x \leq y$ and $y \leq z$, then $x \leq z$ (transitivity)

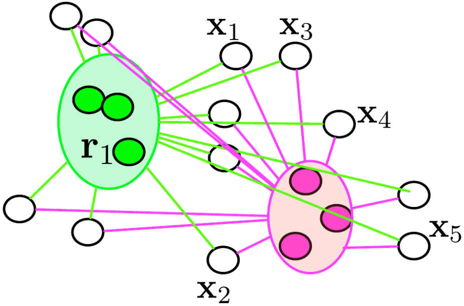
A set X with a partial ordering $\leq$ is called a partial ordered set or **poset**.

Partial Ordering (advantages)

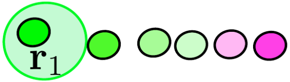
In our representation space a partial ordering has several advantages:

- It does not have to be defined for all pairs of elements of our space.
- It gives a measure of distance with respect to a common comparison point.

Poset Example



$$\mathcal{P}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_5\}$$



Poset Definition

Poset

A Poset P_c with respect to a surrogate class c is the set $\{\dots, x_j, \dots, x_k, \dots\}$ of all unclassified Points x_j, x_k that satisfy the following condition for all $x_i \in C_c$:

$$e^{-||\phi^\theta(x_i)-\phi^\theta(x_j)||} > e^{-||\phi^\theta(x_i)-\phi^\theta(x_k)||} \Leftrightarrow j < k \ \forall\, j, k.$$

Where C_c denotes the points assigned to a surrogate class c. Since elements of C_c are close to each other, compared to other elements, it is enough to represent each class by its medoid.

How to Train the Model?

Our Objective function has to fulfil two goals:

1. Guarantee the classification of elements of C_c as respective surrogate classes.
2. Change the feature space in a way, that pulls samples towards their "near" surrogate classes and away from others.

CNNs objective function

So we need to define a central loss function $\mathcal{L}$ combining two different losses with the mentioned attributes.

$$\mathcal{L}(X, y, R; \theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_1(x_i, y_i; \theta) + \lambda \mathcal{L}_2(x_i, R; \theta)$$

X : Data matrix

y : Surrogate class vector

R : Nearest surrogate classes tensor

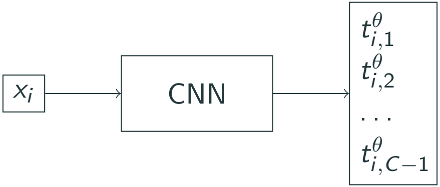
λ : Hyper-parameter for the poset loss

θ : Parameters of the Network (optimization parameter)

Classification Loss

$$\mathcal{L}_1(x_i, y_i; \theta) = -\log \frac{e^{t^{\theta}_{i,y_i}}}{\sum_{j=0}^{C-1} e^{t^{\theta}_{i,j}}} \mathbb{1}_{y_i \neq -1}$$

Where t^{θ}_{i,y_i} represent the logits of sample i given θ.



Poset Loss

$$\mathcal{L}_2(x_i, R; \theta) = -\log \frac{\sum_{z=1}^{Z} e^{\frac{-1}{2\sigma^2}(||\phi^\theta(x_i) - \phi^\theta(r_i^z)||_2^2 - \gamma)}}{\sum_{j=1}^{C'} e^{\frac{-1}{2\sigma^2}(||\phi^\theta(x_i) - \phi^\theta(r_j)||_2^2)}}$$

Z : Number of nearest classes taken into account for every x_i.

C': Number of surrogate classes in the current batch.

γ : Margin between the surrogate classes.

σ : Standard deviation of the current assignment of samples to surrogate classes.

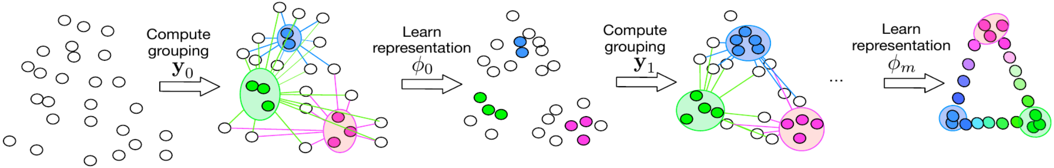
Grouping

There are two interdependent processes, that have to be modelled for training:

- Find a new grouping in the current state of the representation.
- Calculate a new representation, based on new groupings and posets.

For now we have an optimization function for the calculation of a representation with a CNN.

Recap the Method Scheme



Grouping "quality score"

To model the grouping step we construct a function that penalizes large distances inside of a cluster.

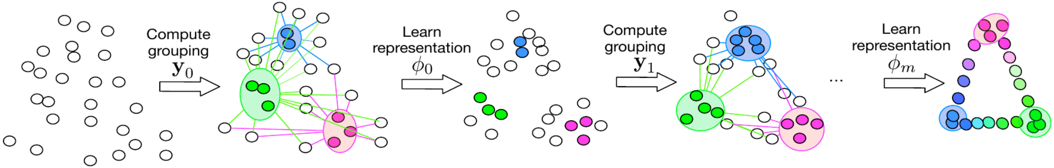
$$\mathcal{G}(X; \phi^{\theta^{m-1}}, y^{(m-1)}) = \sum_{c=0}^{C-1} \frac{\sum_{i:y_i=c} \sum_{j:y_j=c} e^{(-||\phi^{\theta}(x_i) - \phi^{\theta}(x_j)||_2)}}{\left(\sum_{j:y_j=c} 1\right)^2}$$

We maximize this function by the choice of y.

$$y^{(m)} = \underset{y}{\mathbf{argmax}}\ \mathcal{G}(X; \phi^{\theta^{(m-1)}}, y^{(m-1)})$$

$$\textbf{s.t.} \sum_{i:y_i=c} 1 > t,\ \forall c \in \{0, \ldots, C-1\}$$

Summary of the concept



Experiments

Experiments

Evaluation of the method is based on two classical tasks.

- Human pose estimation
- Object recognition

Human pose estimation

Human pose estimation is evaluated with three Datasets:

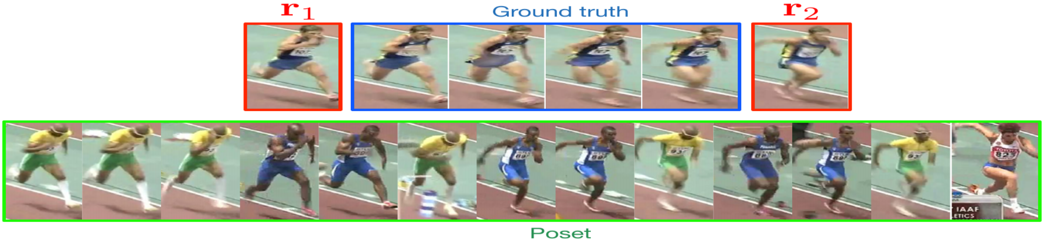
1. Olympic Sports
2. Leeds Sport Pose
3. MPII-Pose

Human pose estimation

Human pose estimation is evaluated with three Datasets:

1. Olympic Sports (zero-shot posture retrieval)
2. Leeds Sport Pose (zero-shot and semi supervised)
3. MPII-Pose (semi supervised posture estimation)

Olympic Sports



AUC Score

AUC score

The AUC score of a classifier is equal to the probability, that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

$$P(\mathrm{score}(x^+) > \mathrm{score}(x^-))$$

AUC Olympic Sports

HOG-LDA [12]	Ex-SVM [16]	Ex-CNN [6]
0.62	0.72	0.64
Alexnet [14]	Doersch et. al [5]	Suffle&Learn [20]
0.65	0.62	0.63
CliqueCNN [3]	Ours scratch	Ours Imagenet
0.83	0.78	**0.85**

Table 1. Avg. AUC for each method on Olympic Sports dataset.

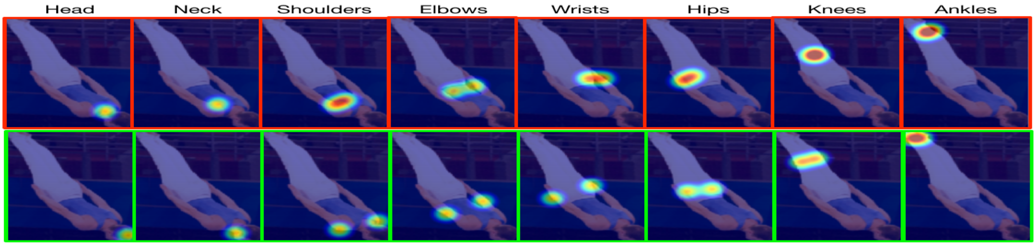
PCP Score Leeps Sport Pose

PCP (percentage of correct parts) means the percentage of correctly classified body parts.

Method	T	UL	LL	UA	LA	H	Total
Ours - Imagenet	83.5	54.0	46.8	34.1	16.8	54.3	**48.3**
CliqueCNN [3]	80.1	50.1	45.7	27.2	12.6	45.5	43.5
Alexnet[14]	76.9	47.8	41.8	26.7	11.2	42.4	41.1
Ours - Scratch	67.0	38.6	34.9	20.5	9.8	35.1	**34.3**
Shuffle&Learn [20]	60.4	33.2	28.9	16.8	7.1	33.8	30.0
Ground Truth	93.7	78.8	74.9	58.7	36.4	72.4	69.2
P. Machines [24]	93.1	83.6	76.8	68.1	42.2	85.4	72.0

Table 2. PCP measure for each method on Leeds Sports dataset for zero-shot pose estimation.

PCK MPII



Thanks for your attention.