# LEARNING FINE-GRAINED IMAGE SIMILARITY WITH DEEP RANKING

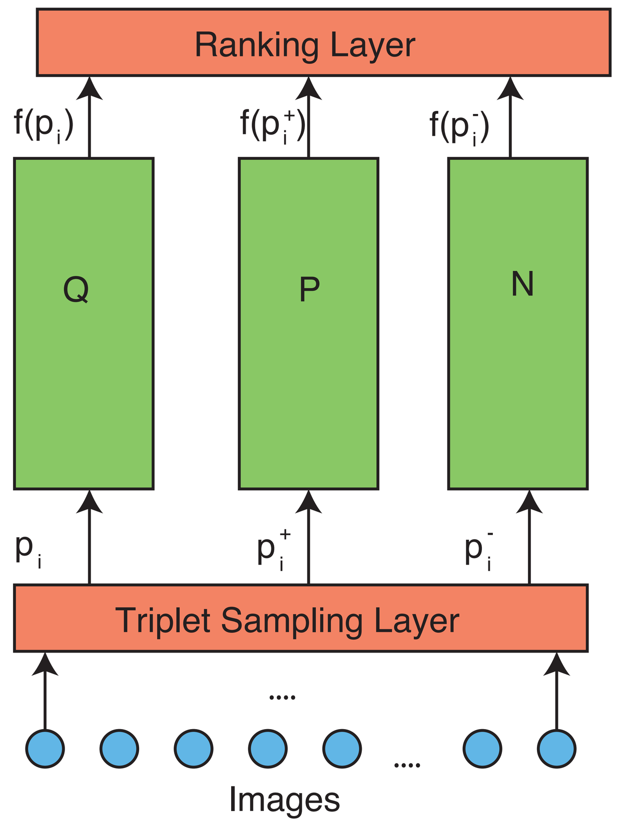
Jiang Wang(Northwestern) Yang Song (Google) Thomas Leung (Google) Chuck Rosenberg (Google) Jingbin Wang (Google) James Philbin (Google) Bo Chen (Caltech) Ying Wu (Northwestern)

## PROBLEM

Fine-grained image similarity, for images with the same category. It is for image-search application, defined by triplets.


- image similarities are defined subtle difference.
- it is more difficult to obtain triplet training data.
- we would like to train a model directly from images instead of rely on the hand-crafted features.

## ARCHITECTURE




- a novel deep learning that can learns fine-grained image similarity model directly from images.
- a multi-scale network structure.
- a computationally efficient online triplet sampling algorithm.
- high quality triplet evaluation dataset.

## DATA

High quality image triplet evaluation dataset:
Available at
https://sites.google.com/site/imagesimilaritydata/

## RELATED WORK

- category-level image similarity: the similarities are purely defined by labels.
- classification deep learning models.
- pairwise ranking model.

## FORMULATION

The similarity of two images $P$ and $Q$ can be defined according to their squared Euclidean distance in the image embedding space:

$$D(f(P), f(Q)) = \| f(P) - f(Q) \|_2^2 \quad (1)$$

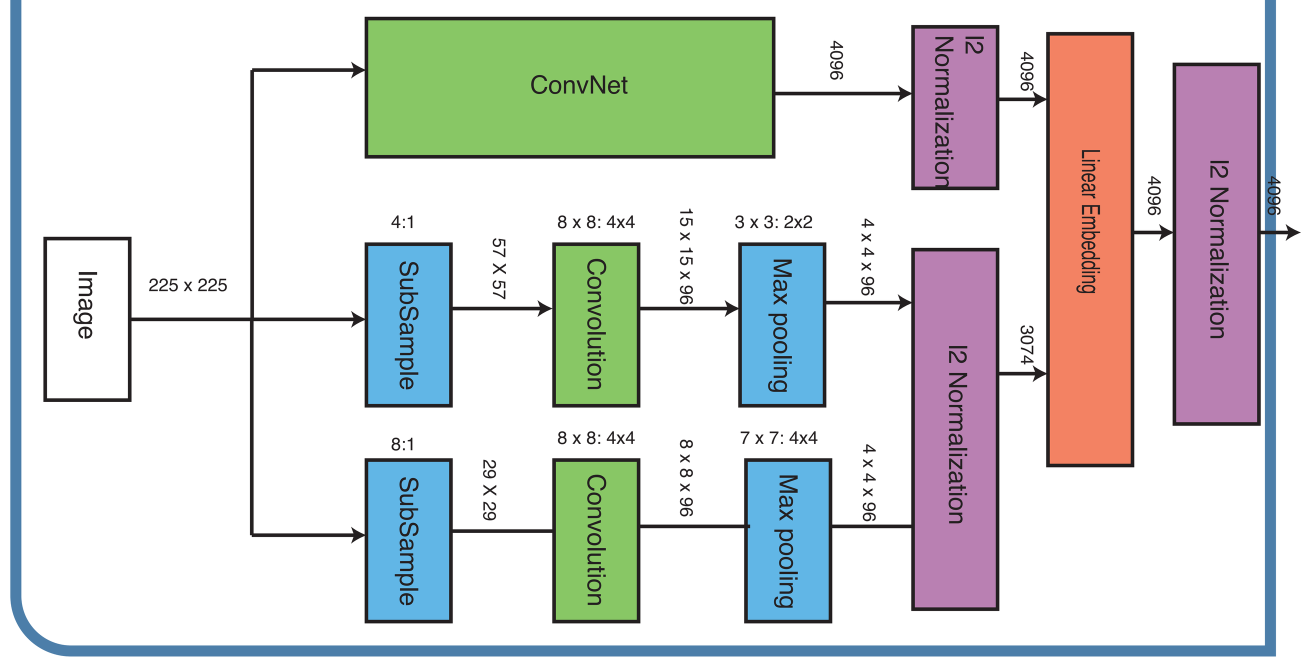
Triplet-based Objective: $r_{i,j} = r(p_i, p_j)$ is pairwise relevance score.

$$D(f(p_i), f(p_i^+)) < D(f(p_i), f(p_i^-)), \quad \forall p_i, p_i^+, p_i^- \text{ such that } r(p_i, p_i^+) > r(p_i, p_i^-) \quad (2)$$

$t_i = (p_i, p_i^+, p_i^-)$ a triplet. The hinge loss is:

$$l(p_i, p_i^+, p_i^-) = \max\{0, g + D(f(p_i), f(p_i^+)) - D(f(p_i), f(p_i^-))\} \quad (3)$$

## MULTI-SCALE ARCHITECTURE




## TRAINING DATA

- ImageNet for pre-training. Category-level information.
- Relevance training data. Fine-grained visual information.
  - Golden Feature, good for visual similarity but not so good for semantic similarity, and it is expensive to compute,

## OPTIMIZATION

- Asynchronized stochastic gradient algorithm.
- Momentum algorithm.
- Dropout to avoid overfitting

Challenges:

- Cannot enumerate all the triplets, need to sample important triplets.
- Cannot load all the images into memory, need to generate triplets online.

## TRIPLET SAMPLING

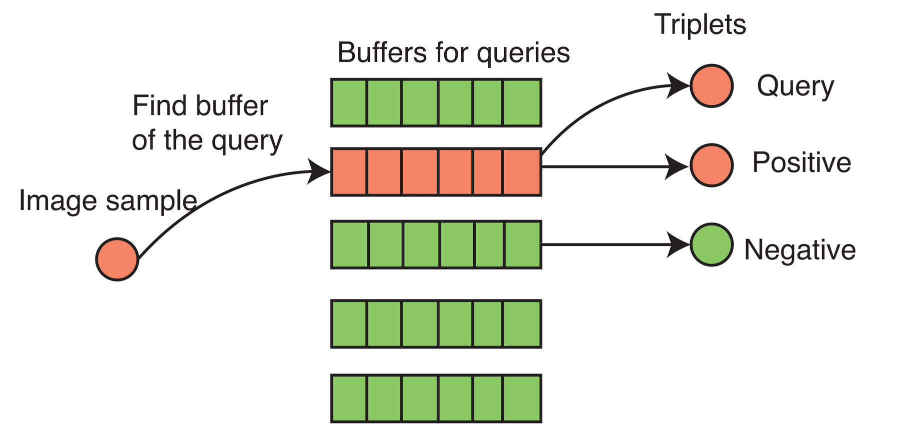
Sampling criteria: we sample more highly relevant images.

Total relevance score $r_i$:

$$r_i = \sum_{j: c_j = c_i, j \neq i} r_{i,j} \quad (4)$$

- For query image: according to total relevance score.
- For positive image: sample images with the same label as the query image, sampling probability is $P(p_i^+) = \frac{\min\{T_p, r_{i,i^+}\}}{Z_i}$.
- For negative image, we have two types of samples:
  1. in-class negative: we draw in-class negative samples $p_i^-$ with the same distribution as the positive image. We also require that the margin between the relevance score $r_{i,i^+}$ and $r_{i,i^-}$ should be larger than $T_r$
  2. out-of-class negative: drawn uniformly from all the images in different categories.

Online triplet sampling: reservoir sampling:




## EXPERIMENTS

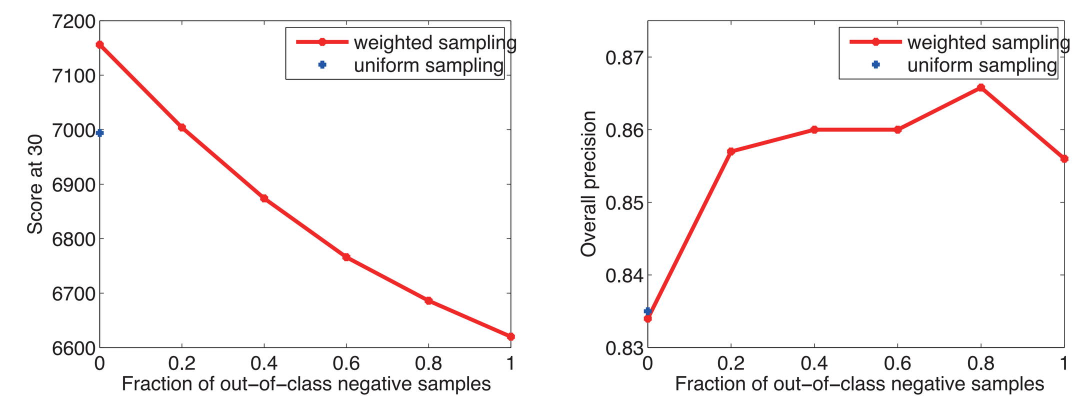
Comparison with hand-crafted features:

| Method | Precision | Score-30 |
|---|---|---|
| Wavelet | 62.2% | 2735 |
| Color | 62.3% | 2935 |
| SIFT-like | 65.5% | 2863 |
| Fisher | 67.2% | 3064 |
| HOG | 68.4% | 3099 |
| SPMKtexton1024max | 66.5% | 3556 |
| L1HashKPCA | 76.2% | 6356 |
| OASIS | 79.2% | 6813 |
| Golden Features | 80.3% | 7165 |
| DeepRanking | 85.7% | 7004 |

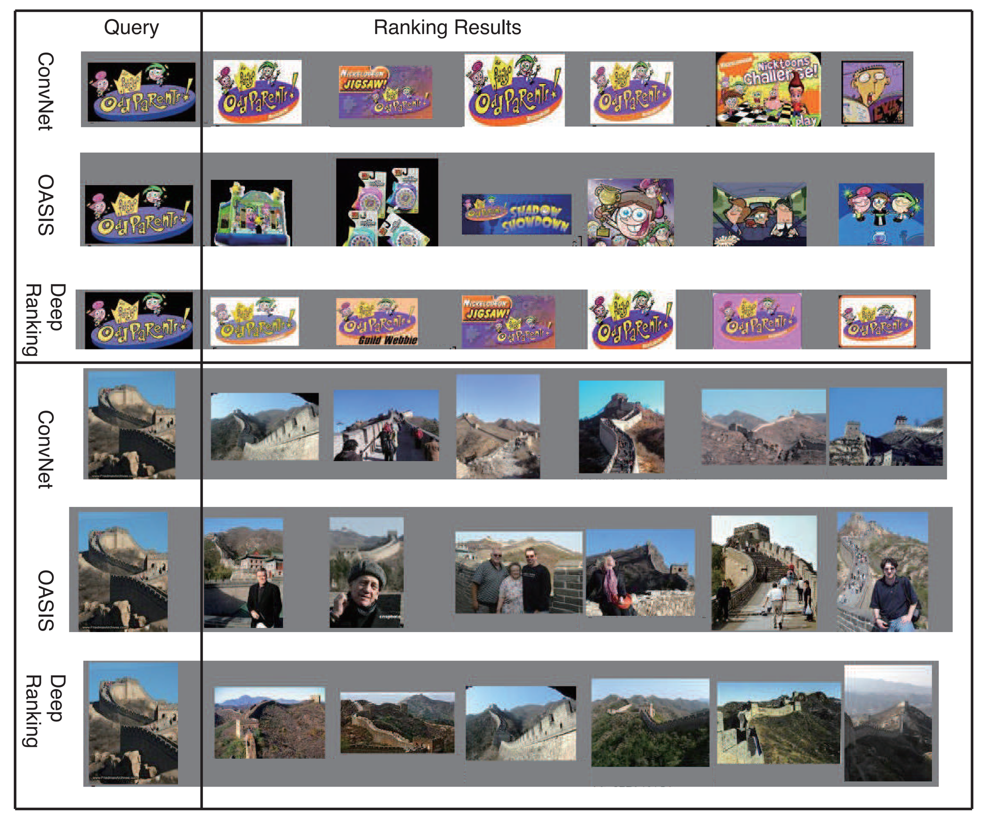
Comparison of different architectures:

| Method | Precision | Score-30 |
|---|---|---|
| ConvNet | 82.8% | 5772 |
| Single-scale Ranking | 84.6% | 6245 |
| OASIS on Single-scale Ranking | 82.5% | 6263 |
| Single-Scale & Visual Feature | 84.1% | 6765 |
| DeepRanking | 85.7% | 7004 |

Comparison of different sampling methods:




## RANKING EXAMPLES




## ACKNOWLEDGMENT

The work was done when the first author is working as an intern at Google.