Deep Fine-grained Vehicle Recognition from Surveillance Imagery

Luke Jaffe and Wesam Sakla
Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Overview

- We want to distinguish between vehicles detected in surveillance imagery.

Image @ time t_1

Image @ time t_2

=

?

Vehicle Retrieval

Image Representations

Image Matching

Ranked List

Probe Image

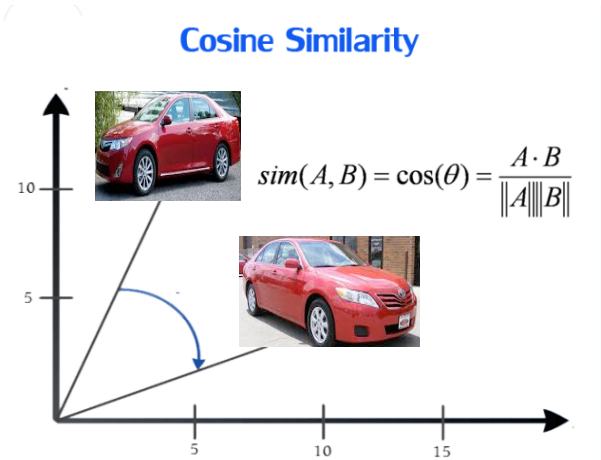
$v = (v_1, \ldots, v_n)$

Gallery Images

$v_1 = (v_{11}, \ldots, v_{1n})$

$\vdots$

$v_k = (v_{k1}, \ldots, v_{kn})$



Similarity Score	Image
0.97	
0.95	
⋮	⋮
0.12	
0.04	

We will use deep convolutional neural networks to learn robust image representations.

Data

- Vehicle ID (VID) dataset: CVPR 2016 paper*
- Largest instance-level dataset of vehicles
- Color images acquired by multiple surveillance cameras in China
  - varying spatial resolutions, various backgrounds, variety in illumination
  - 26,328 unique vehicle instances (at least 2 images/instance, average of ~8 images/instance)
  - Instance annotations generated using license plate information
  - High resolution, oblique overhead viewing geometry (front, rear)

Partition	# Vehicles	# Images
$Train$	13164	113346
$Test$	13164	108221
TOTAL	***26328***	***221567***

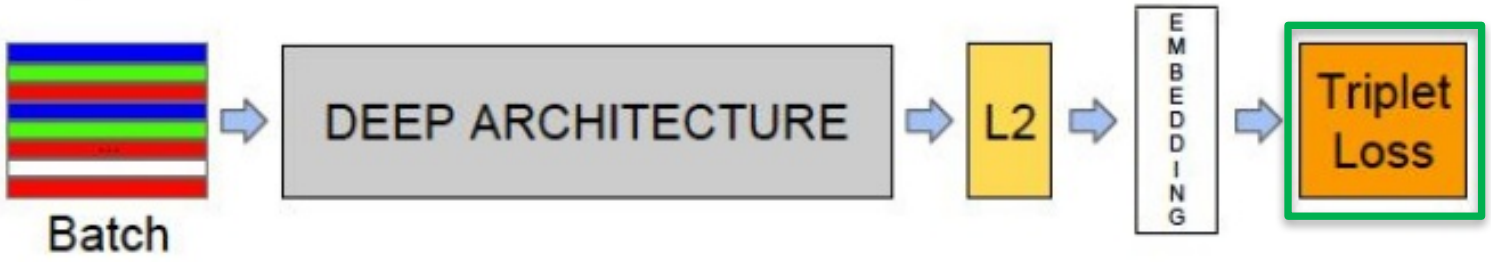
*Liu *et al.*, *Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles*, at CVPR 2016.

Data

*Liu *et al.*, *Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles*, at CVPR 2016.

Fine-grained Vehicle Recognition

- Utilize ideas found in *FaceNet** *(Google) for facial recognition*
- *Train a CNN to learn a compact feature space (128D) representation where distances directly correspond to a measure of vehicle similarity*

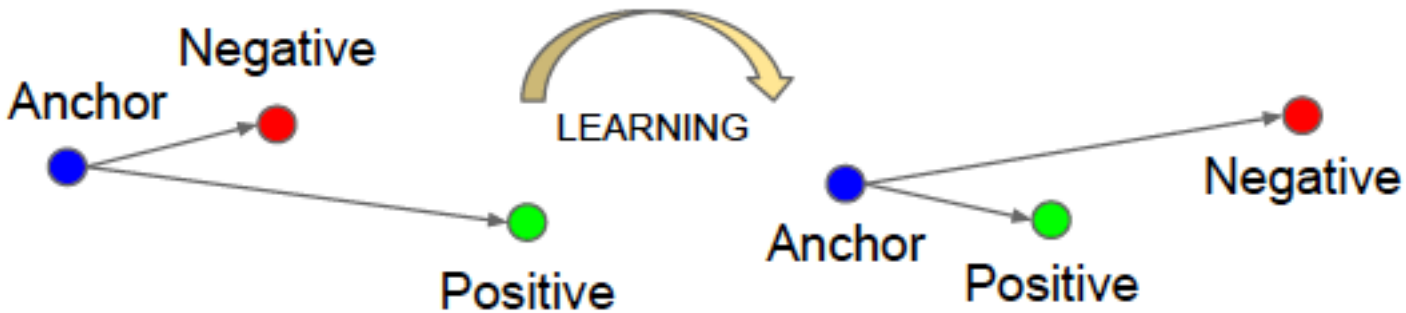


What is a Triplet?

*Schroff *et al.*, *FaceNet: A Unified Embedding for Face Recognition and Clustering*, at CVPR 2015.

Triplets

- During training, the network is presented with batches of cleverly sampled *triplets*.
- A triplet is a selection of three images $< x_i^a, x_i^p, x_i^n >$ for a given vehicle instance i.
  - x_i^a: *Anchor*
  - x_i^p: Positive
  - x_i^n: Negative

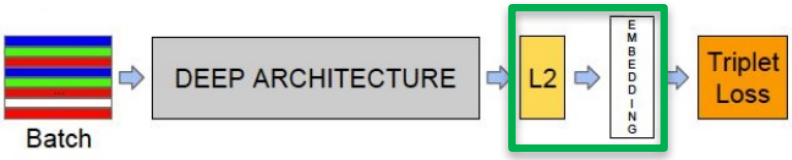


$$\left\|(x_i^a) - f(x_i^p)\right\|_2^2 + \alpha < \|(x_i^a) - f(x_i^n)\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in T$$

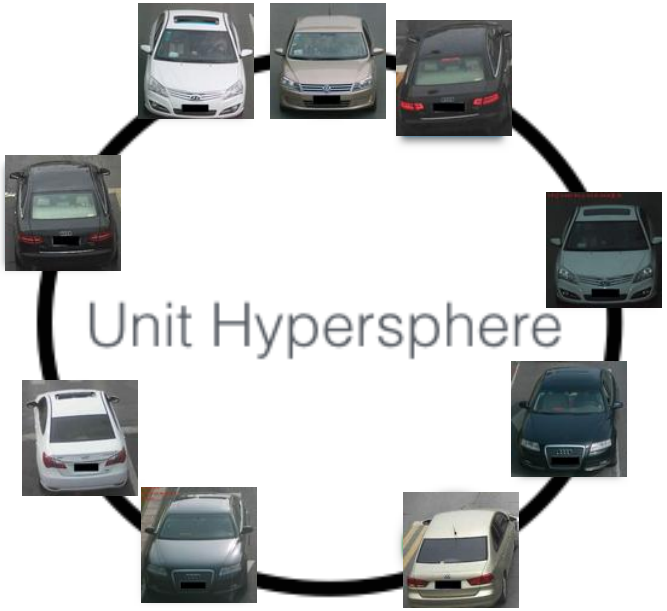
Triplet Loss: $\sum_i^N \left[\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+$ where f is the embedding

*Schroff *et al.*, *FaceNet: A Unified Embedding for Face Recognition and Clustering*, at CVPR 2015.

Semantic Embeddings

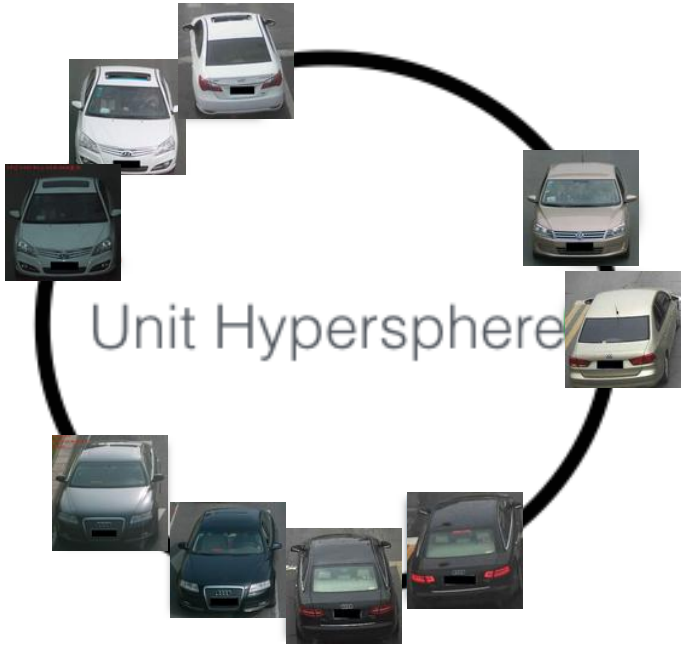


Vehicle Instances



No Semantic Organization

Training

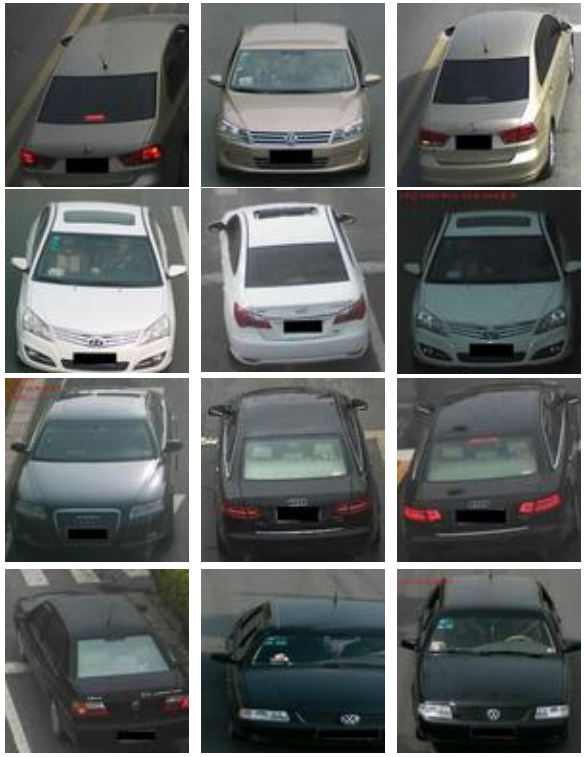


Vehicles Distinguishable

Images of same vehicle are close; images of different vehicles are farther apart.

Mini-batch Triplet Sampling

- *P* images per vehicle
- *Q* vehicles from dataset
- *P = 3, Q = 4*



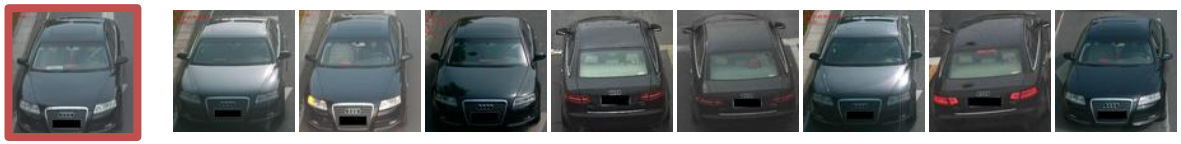
Generate all anchor-positive combinations: $N = Q\binom{P}{2}$

$$N = 4\binom{3}{2} = 12$$

Randomly select negatives to generate N triplets

Vehicle Retrieval

- Asks the question: "*Given a new image of a vehicle, how reliably can we find previously seen instances of it?*"



	Small	Medium	Large	XL	XXL	FULL
$\#\ of\ vehicles$	800	1600	2400	3200	6000	13164
$\#\ of\ images$	6493	13377	19777	26353	48922	108221

Test Data Splits for VID Dataset.

- Suppose we have N_i images for vehicle i in a given test set.
- Randomly place $N_i - 1$ images for vehicle i into *gallery set*.
- Place remaining image of each vehicle into *probe set*.
- Evaluation metric: *mean average precision (MAP)*

Mean Average Precision (MAP)

Probe (query)

Gallery instances

Ranking #1

Recall	0.20	0.40	0.40	0.60	0.60	0.60	0.80	1.0
Precision	1.0	1.0	0.67	0.75	0.60	0.50	0.57	0.625

Ranking #2

Recall	0.20	0.40	0.60	0.60	0.60	0.80	1.0
Precision	1.0	1.0	1.0	0.75	0.60	0.67	0.714

AP for Ranking #1: (1.0 + 1.0 + 0.75 + 0.57 + 0.625)/5 = 0.789

AP for Ranking #2: (1.0 + 1.0 + 1.0 + 0.67 + 0.714)/5 = 0.8768

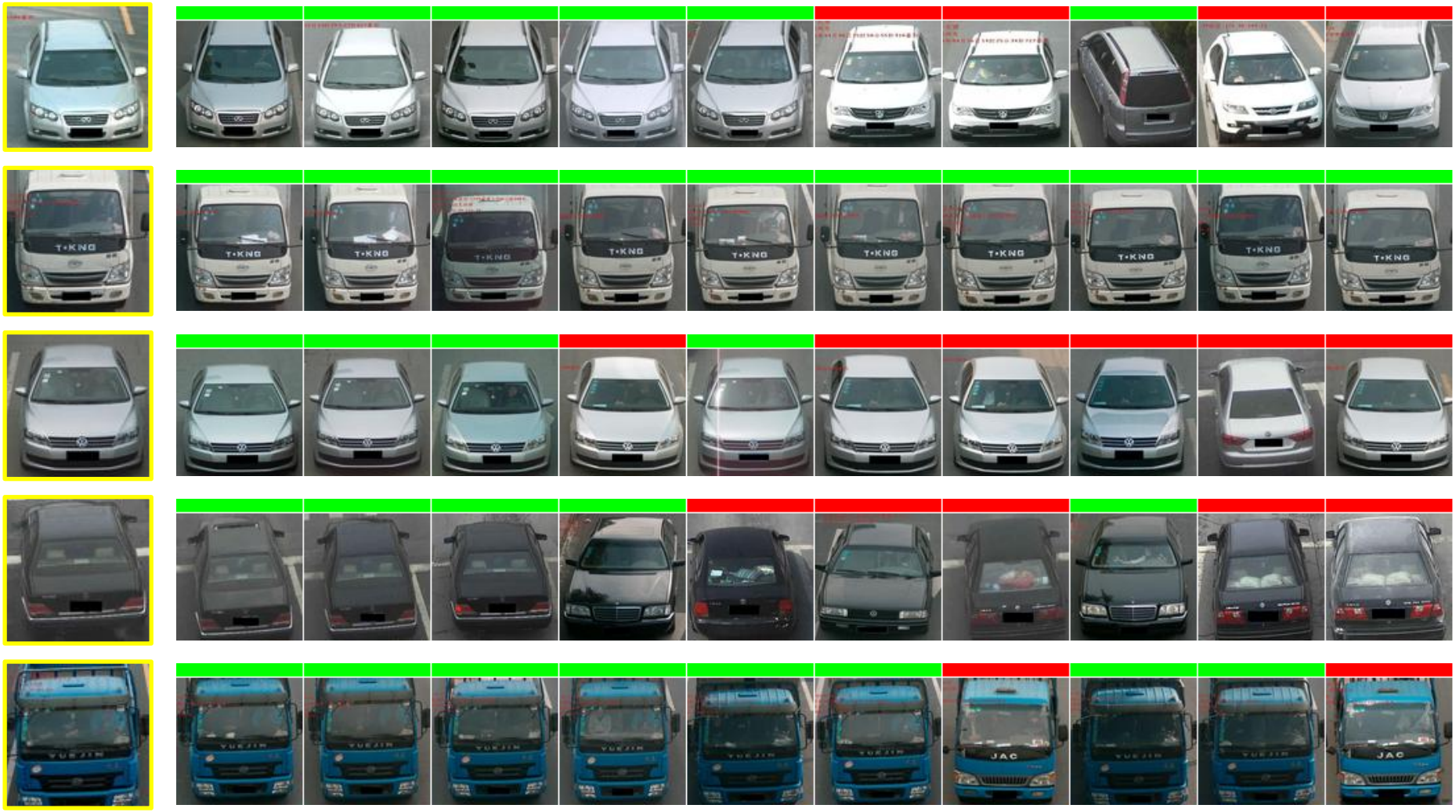
Quantitative Results

Algorithm	Spatial Resolution (pixels)	Small (800)	Medium (1600)	Large (2400)
*VGG + Coupled Cluster Loss**	$224x224$	0.492	0.448	0.386
*Mixed Diff + Coupled Cluster Loss**	$224x224$	0.546	0.481	0.455
Ours - Inception	$224x224$	0.7771	0.6848	0.6366
Ours - Inception	$96x96$	0.7509	0.6638	0.6066
Ours - Inception	$48x48$	0.6927	0.5966	0.5393
Ours – ResNeXt	$32x32$	0.6461	0.5432	0.4896

MAP of Retrieval Task as a Function of Image Resolution and Gallery Size with $n_{trials} = 10$.

*Liu *et al.*, *Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles*, at CVPR 2016.

Qualitative Results ($96x96$)



Questions?

Lawrence Livermore
National Laboratory