Unsupervised deep clustering for semantic object retrieval

Steven Hickson, Anelia Angelova, Irfan Essa, Rahul Sukthankar

Google Brain

Research at Google

Motivation

Observe motion and extract moving agents.

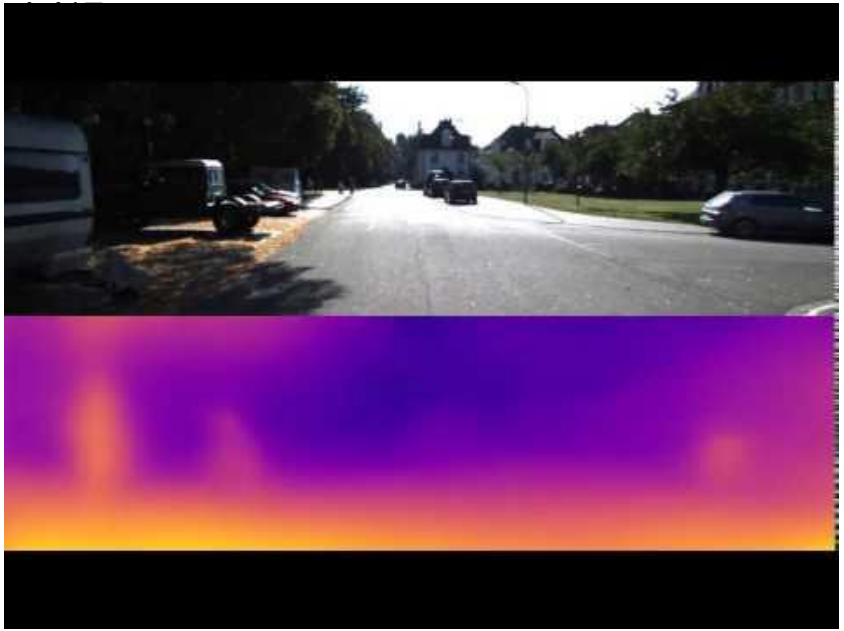
These must be entities. i.e., full objects.

Unsupervised object discovery to form semantic classes of objects.

Video credit Tinghui Zhou: https://people.eecs.berkeley.edu/~tinghuiz/projects/SfMLearner/

We (almost) know how to do SFM (with deep nets)

SFMLearner: T. Zhou et al. ’17



Unsupervised learning of depth and egomotion

https://people.eecs.berkeley.edu/~tinghuiz/projects/SfMLearner/
https://youtu.be/RTFatijYcaU

SFMNet: S. Vijayanarasimhan et

Additionally, learning of motion masks.

Main idea

You can extract moving objects which will be entities.

We won’t know their class but will discover semantic affiliation.

The goal is to (learn to) detect them in out-of-sample images.

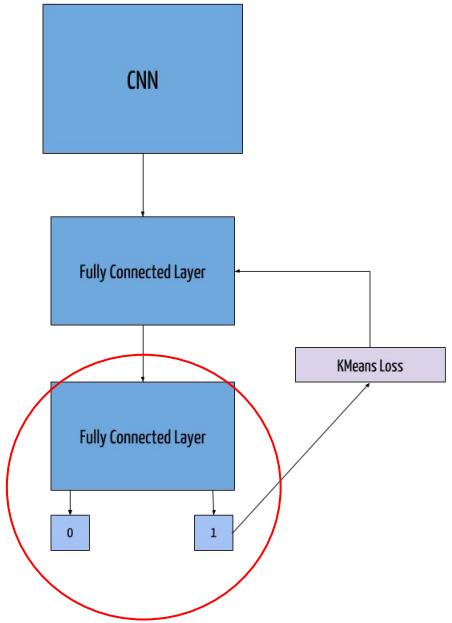
Unsupervised!

Clearly all these apply to weakly supervised
or semi-supervised tasks.

This work

Moving objects can be used to form an embedding.

Learn an object vs background discriminator.

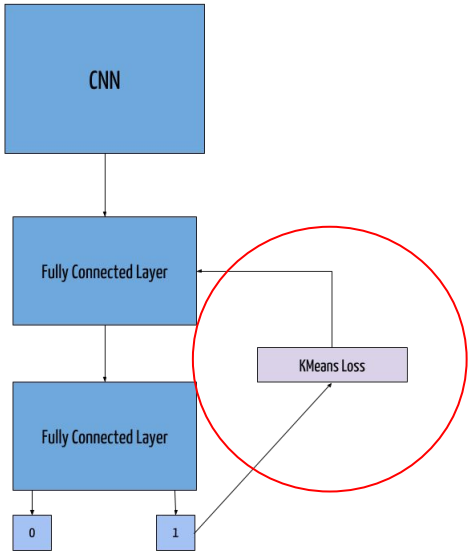


This work

Moving objects can be used to form an embedding.

Learn: object vs background

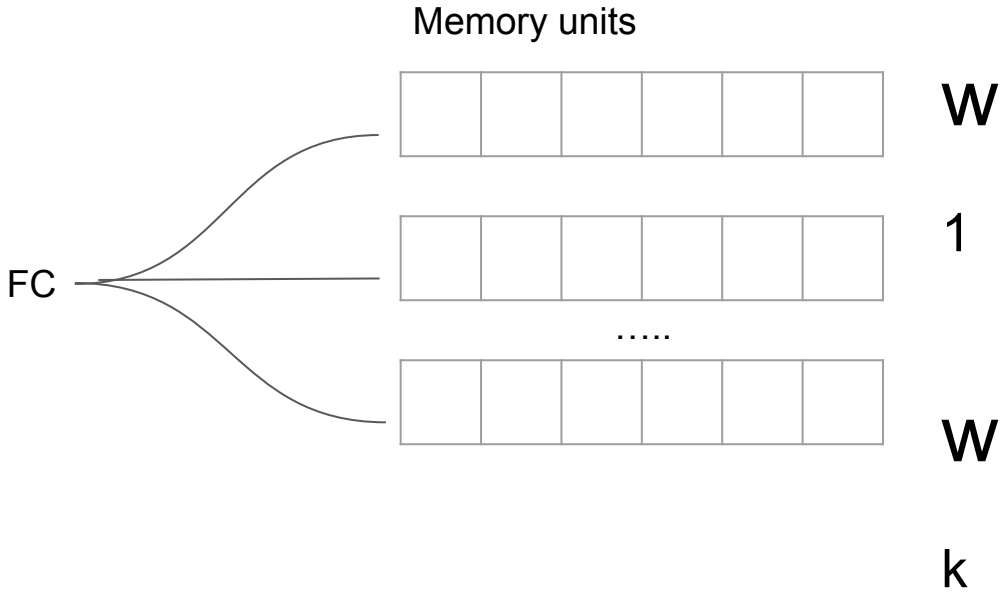
Improve embedding by forcing objects to cluster.



Differential clustering to improve embedding

Clustering objective

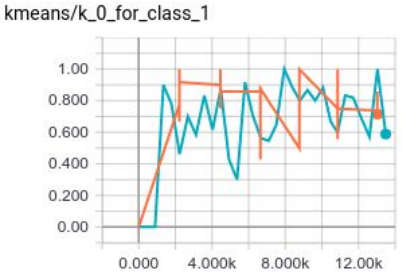
$$L_K=\Sigma_n \min_k[(x_n-w_k)^2]$$

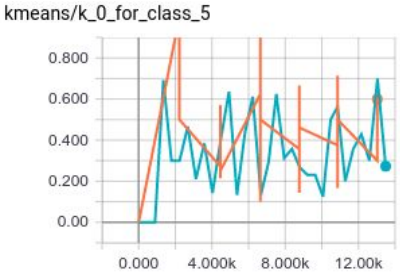


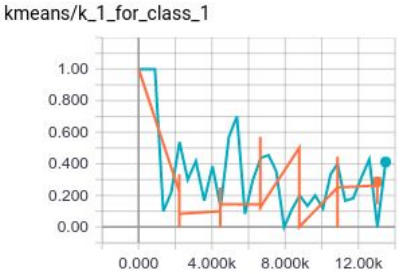
$$\text{Min } L= L_K+\alpha L_2+\beta L_C$$

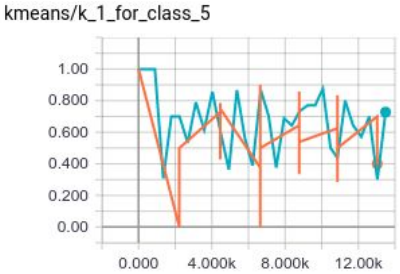
With additional L_2 regularization and L_c is loss balancing the size of the clusters

Experiments: Cifar









Two classes from Cifar 10
Evaluation process uses the labels for visualization (above). The figures show accuracy per learned cluster as a function of time.

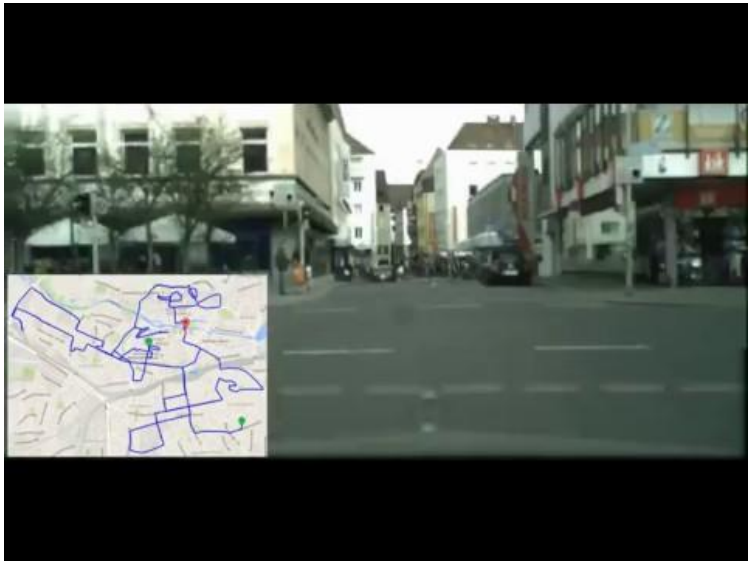
	Class dog	Class auto
Cluster 0	68.5%	17.9%
Cluster 1	31.5%	82.1%

We also tried contrastive loss : Hadsell et al.Since the task is hard, no obvious clusters were formed.

Experiments: The Cityscapes data

Segmentation masks provided for 1/30s of the data.

We use them here but idea is to use all unsupervised data.



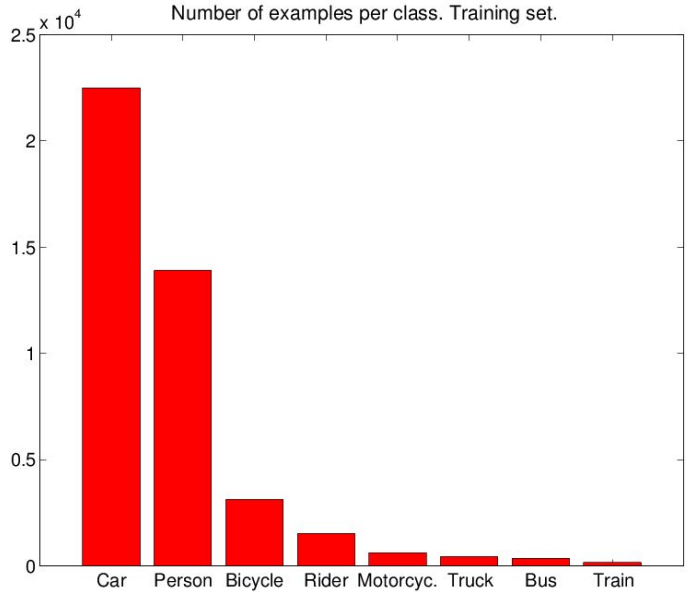
From:https://www.cityscapes-dataset.com/examples/

Retrieval results: Cityscapes data

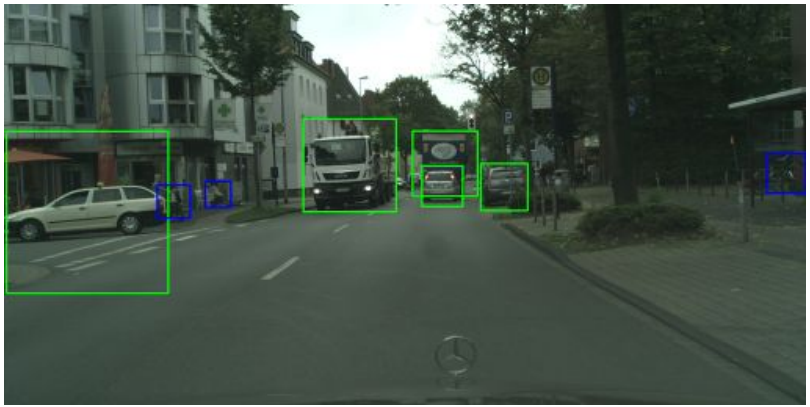
Training: build foreground/background and clustering objective embedding

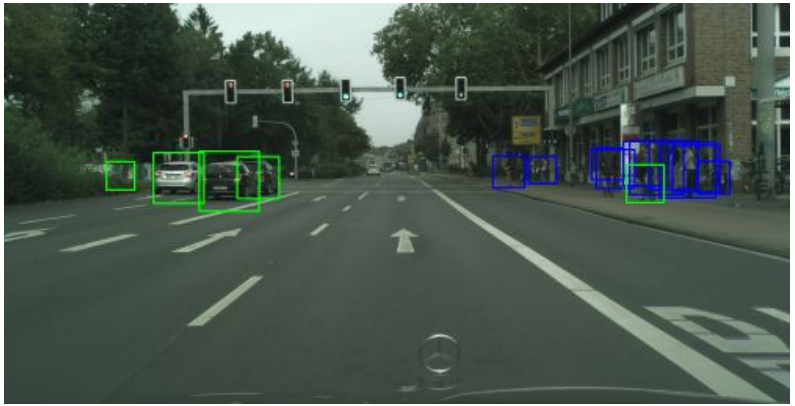
Testing: cluster into several groups (known annotation for eval only)

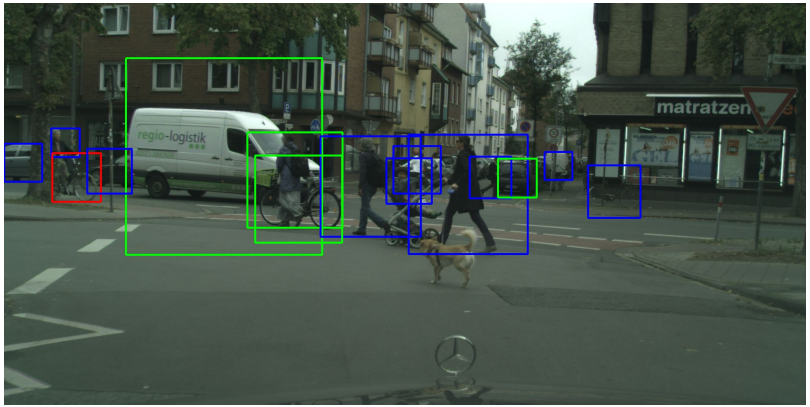
Large imbalance of data.
Data is also quite noisy.

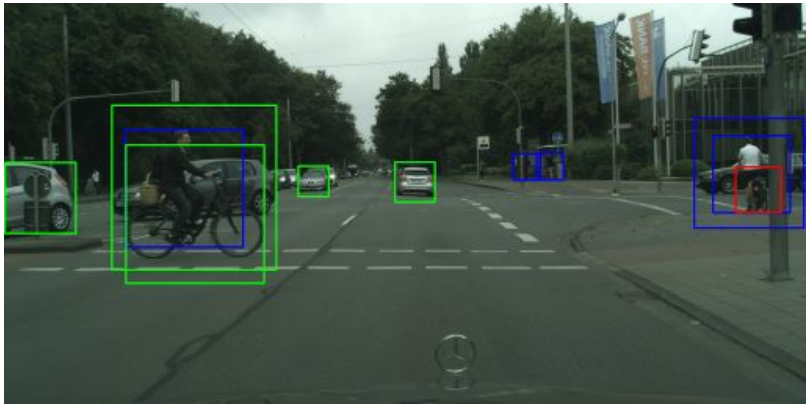


Retrieval results: Cityscapes data



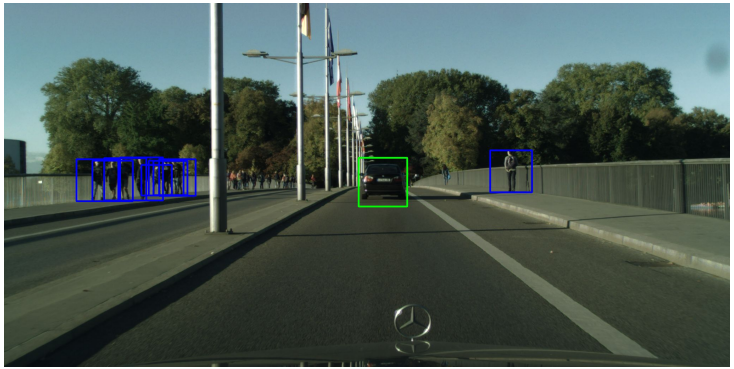


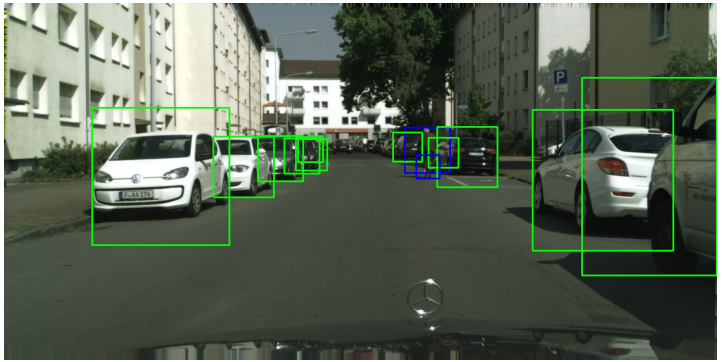


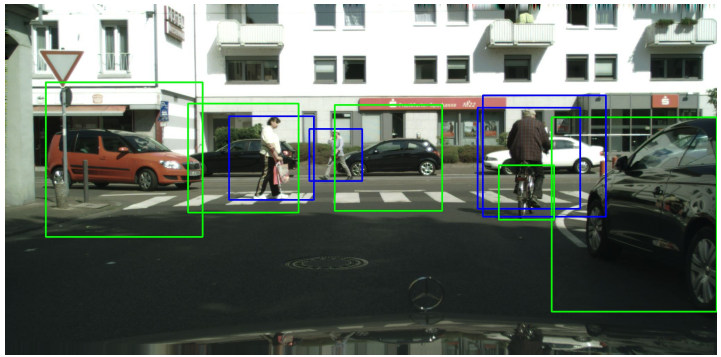


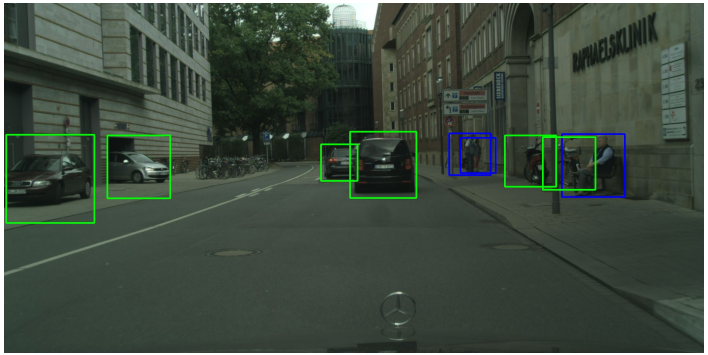
Class 1
Class 2
Class 3

Retrieval results: Cityscapes data









Class 1
Class 2
Class 3

Note: since data is very noisy, it is really hard to form clustering.
E.g a bicycle may have a car in the background. A bicycle is likely to have a person on it.

Clustering results

Classes	Cluster 0	Cluster 1	Cluster 2
Person	4302	198	29
Rider	634	161	17
Car	690	5053	538
Truck	34	92	10
Bus	25	117	7
Train	12	23	3
Motorcycle	73	119	21
Bicycle	583	946	180

Comparison to the baseline embedding (i.e. when discriminating background vs object):

Classes	Cluster 0	Cluster 1	Cluster 2
Person	4428	1	0
Rider	813	0	0
Car	6292	13	0
Truck	126	2	0
Bus	152	0	0
Train	33	0	0
Motorcycle	205	0	0
Bicycle	1698	0	0

Classification accuracy 66%-69% when considering the 3 main classes: person, car,bicycle

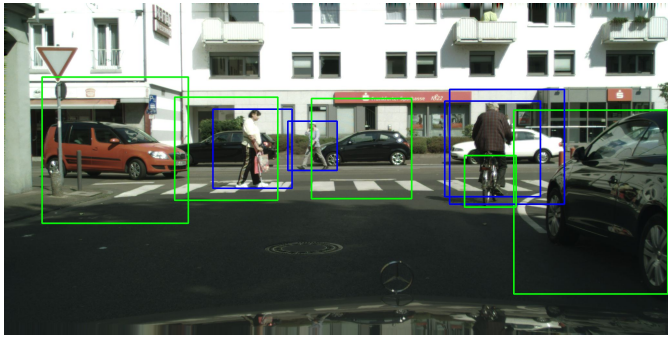
Summary

Can we retrieve semantically related objects from videos?

Clustering is implemented in a DNN with memory units.

Experiments with Cityscapes dataset for moving objects.

Retrieval of meaningful classes.



Future: This is a very challenging task (class overlap)

Base embedding is also based on noisy data

Suggestions for datasets/embeddings, where to try the approach.

Thank you!
Questions?

anelia@google.com