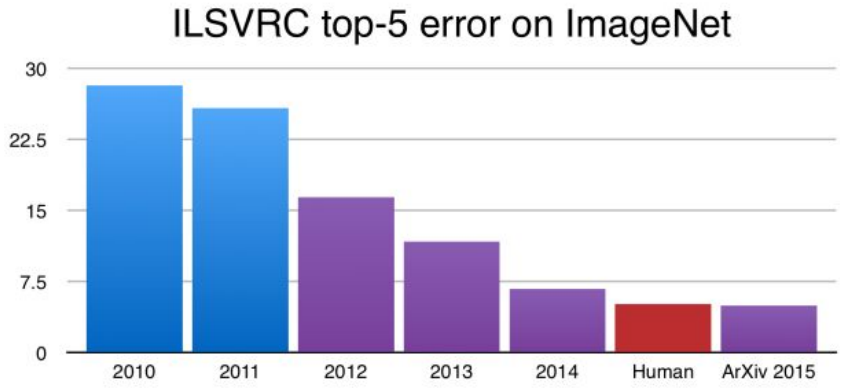
# Unsupervised learning of visual representations using videos

X. Wang and A. Gupta
ICCV 2015

Experiment presentation by Ashish Bora

# Motivation

- Supervised methods work very well




- But labels are expensive
- Lot of unlabeled data is available
- Can we learn from this huge resource of unlabeled data?

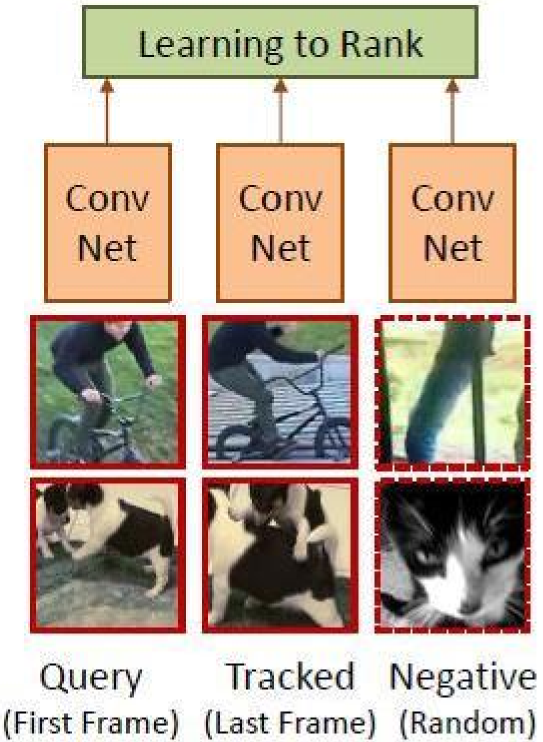
Image from : https://devblogs.nvidia.com/wp-content/uploads/2015/08/image1-624x293.png

# Approach

- Learn a vector representation for image patches in a video
  - Similar patches should be close (cosine similarity)
  - Random patches should be far
- Ranking Loss

$$\min_W \frac{\lambda}{2} \| W \|_2^2 + \sum_{i=1}^{N} \max\{0, D(X_i, X_i^+) - D(X_i, X_i^-) + M\}$$

- CNN architecture similar to AlexNet



(b) Siamese-triplet Network

Image from : http://www.cs.cmu.edu/~xiaolonw/unsupervise.html

# How to get patches?

Positive pairs

- Tracking across time provides self-supervision
- Get the bounding box for first image using SURF with Improved Dense Trajectories.

Negative Pairs

- Random sampling
- Hard-negatives for better training

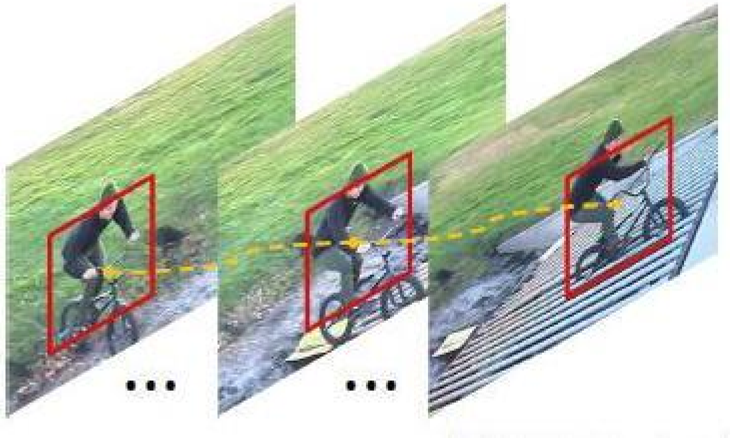


Image from : http://www.cs.cmu.edu/~xiaolonw/unsupervise.html

# Experiments - Outline

- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# Experiments - Outline

- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# tSNE - a quick introduction

- tSNE = t-Distributed Stochastic Neighbor Embedding
- Want to visualize a set of data-points in n-dimensional space
- Visualization beyond 3-D is hard
- tSNE: A method to embed each datapoint to small number of dimensions (2 or 3) such that small/local distances are preserved
- Contrast: PCA preserves large distances
- For more details, see: https://www.youtube.com/watch?v=RJVL80Gg3lA

# tSNE on hw2 images

- Color similarity
- Backgrounds
- Black and white images

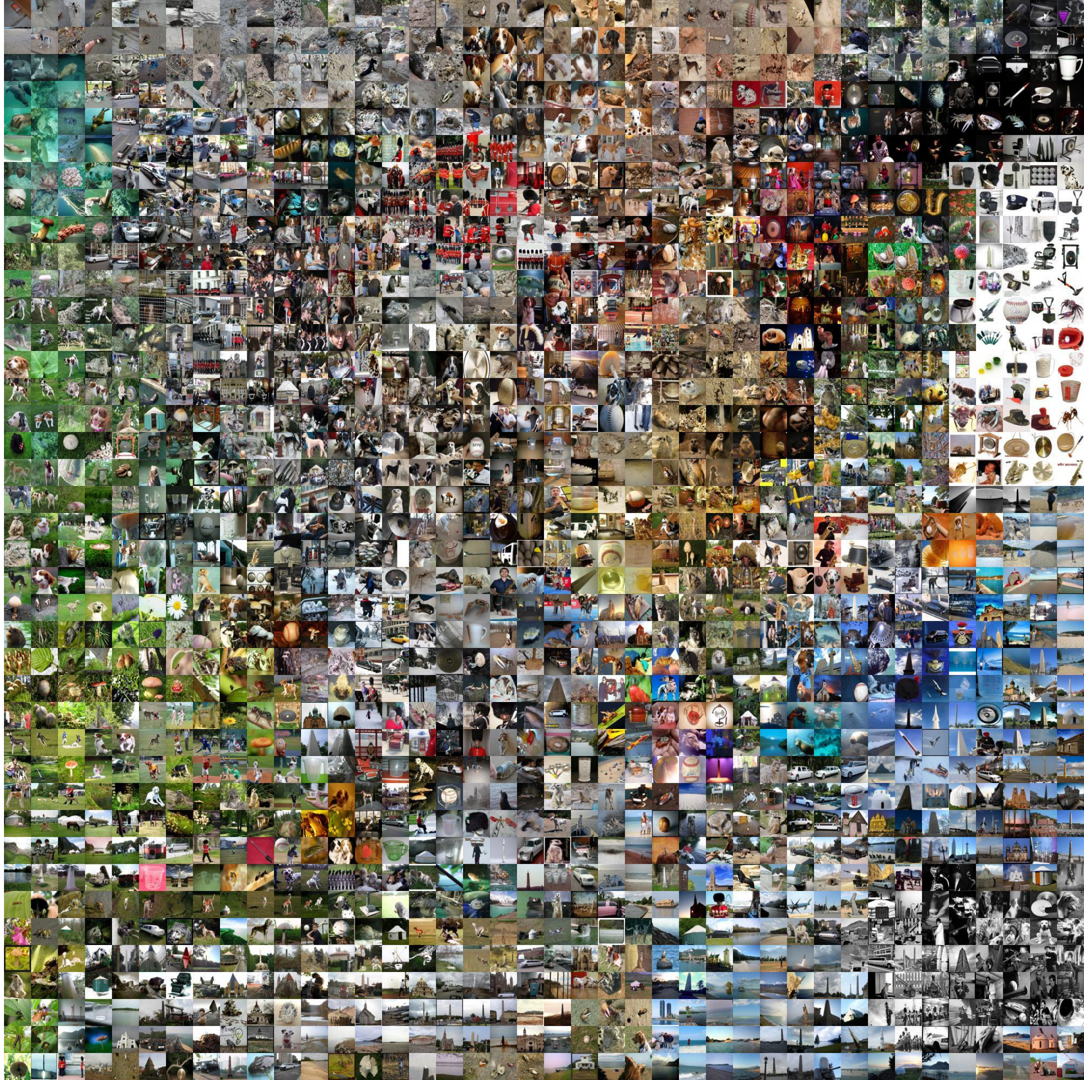


Image generated with code from : http://cs.stanford.edu/people/karpathy/cnnembed/

# tSNE Results

# tSNE Results

# tSNE Results

# tSNE on Stanford40

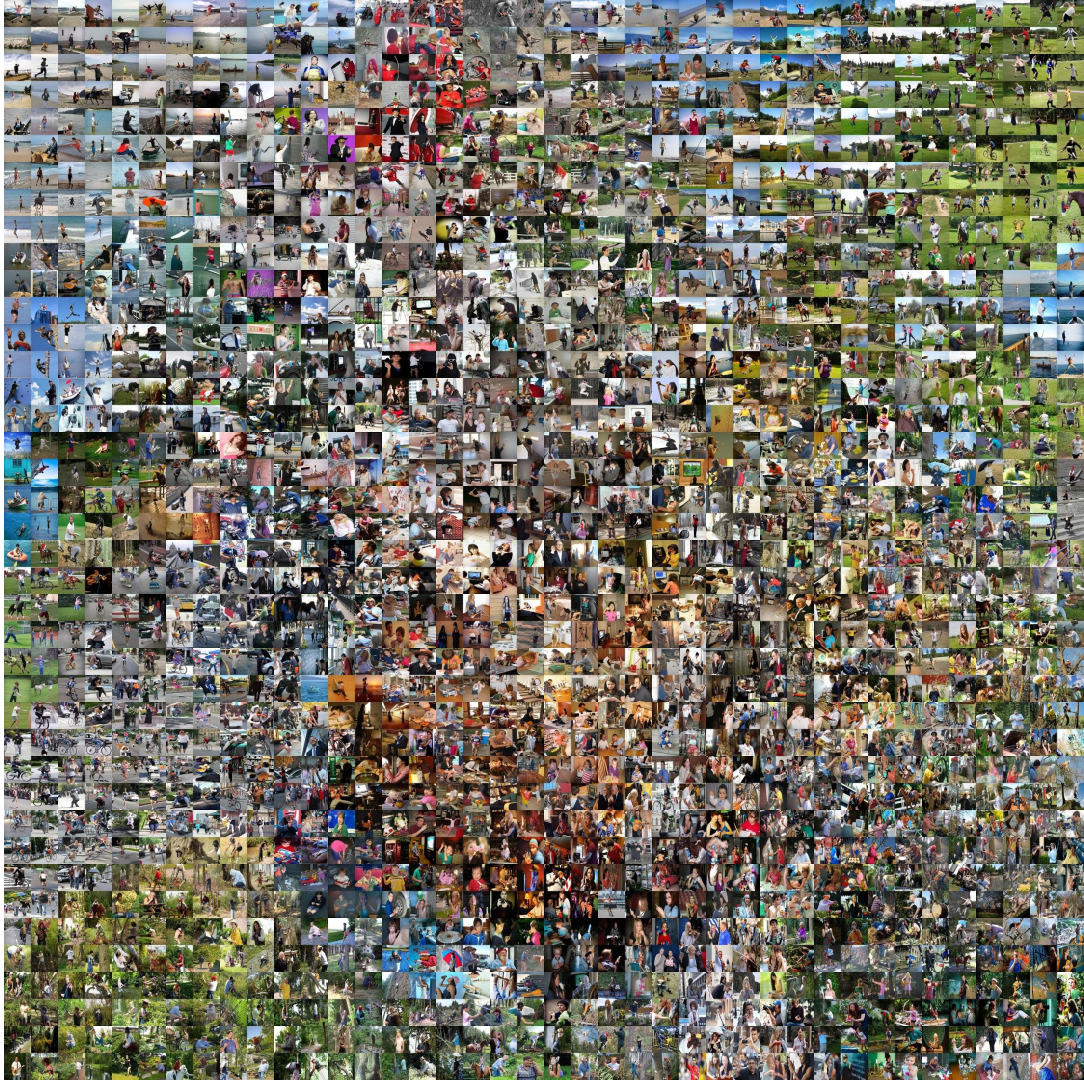
- Learned from videos
- Do we get clusters specific to activities?

Results

- Most clusters are based on background and objects (bikes, boats) rather than activity

http://vision.stanford.edu/Datasets/40actions.html

Image generated with code from : http://cs.stanford.edu/people/karpathy/cnnembed/



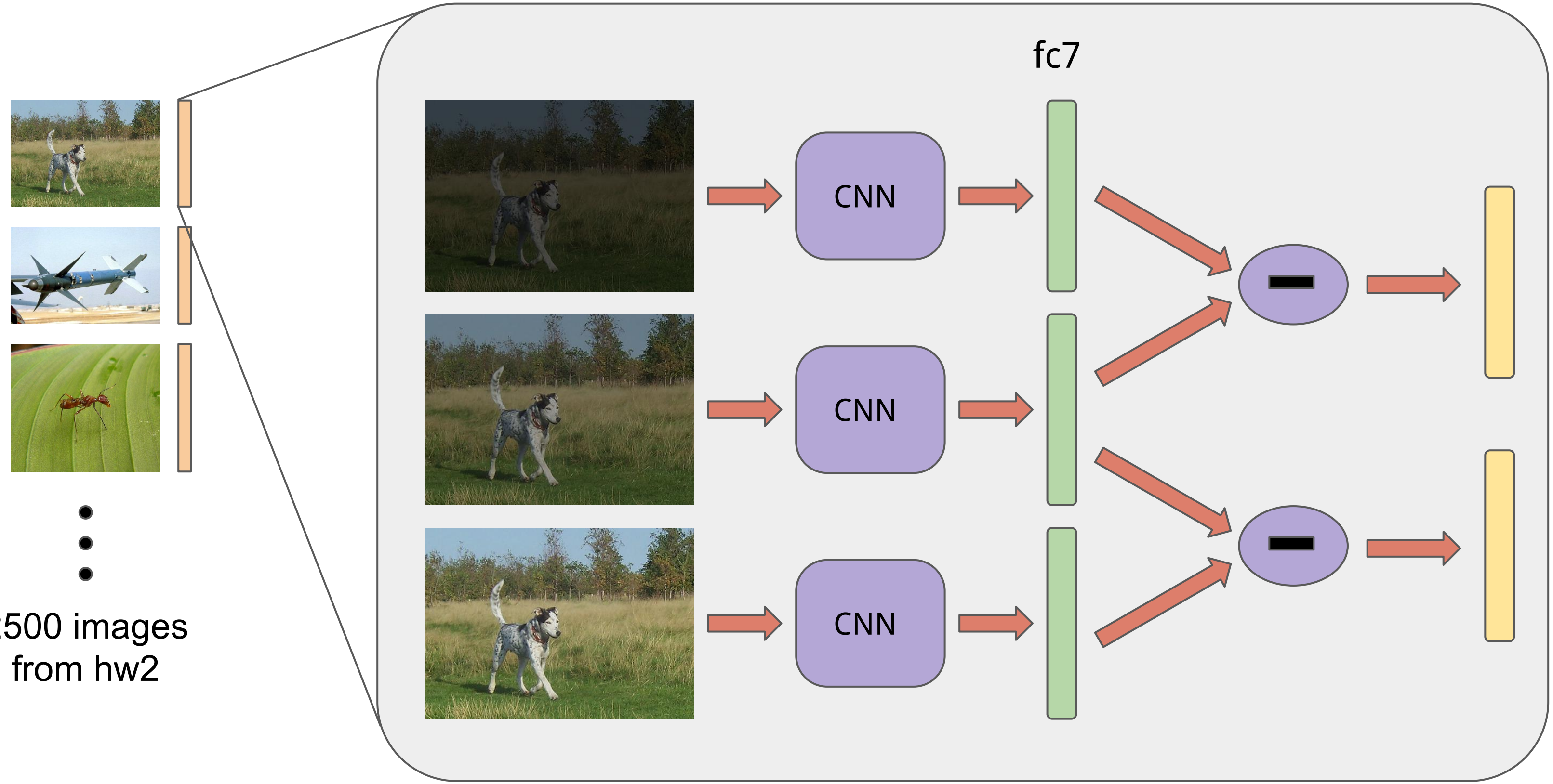
# Experiments - Outline

- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# Input variation

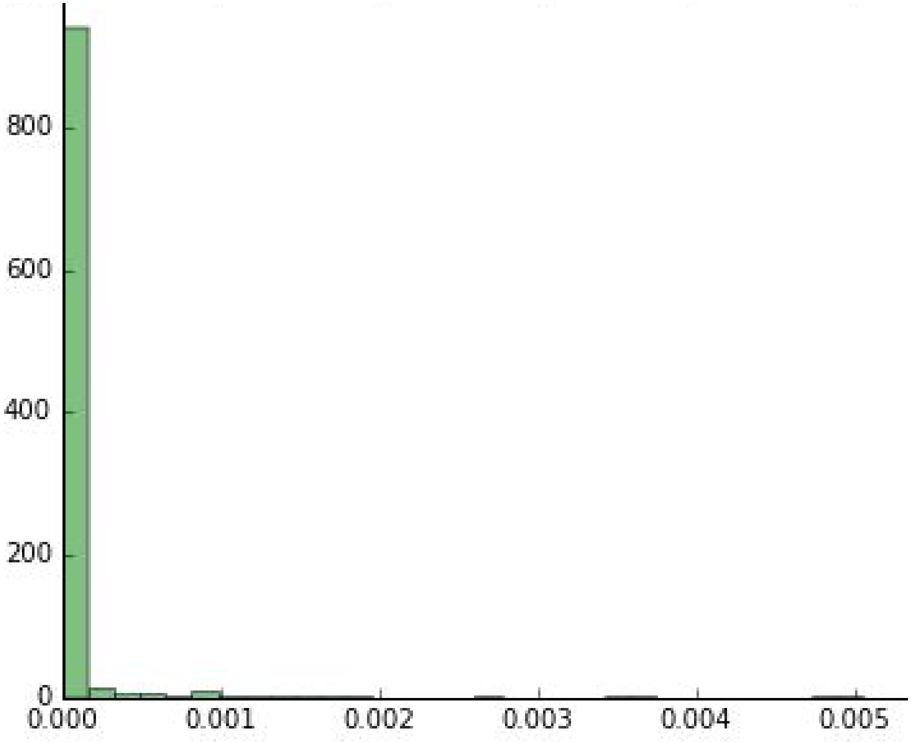
- Input is 227 x 227, but output is only 1024 dimensional
- Some things must be thrown away
- Illumination, saturation, rotation unimportant to recognize images that co-occur, which is the objective for unsupervised phase.
- Verify that these invariances are learned

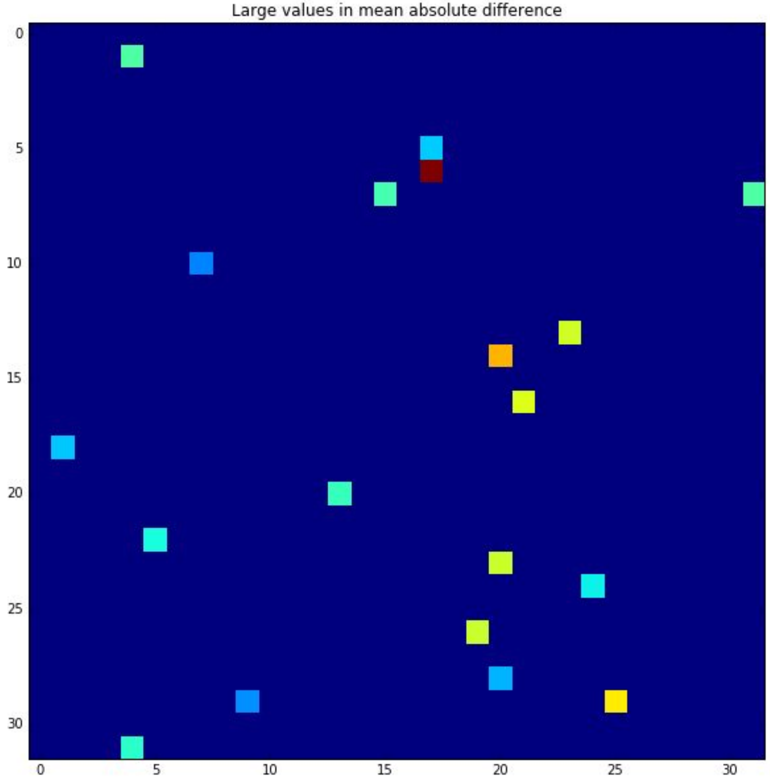
# Input variation - illumination



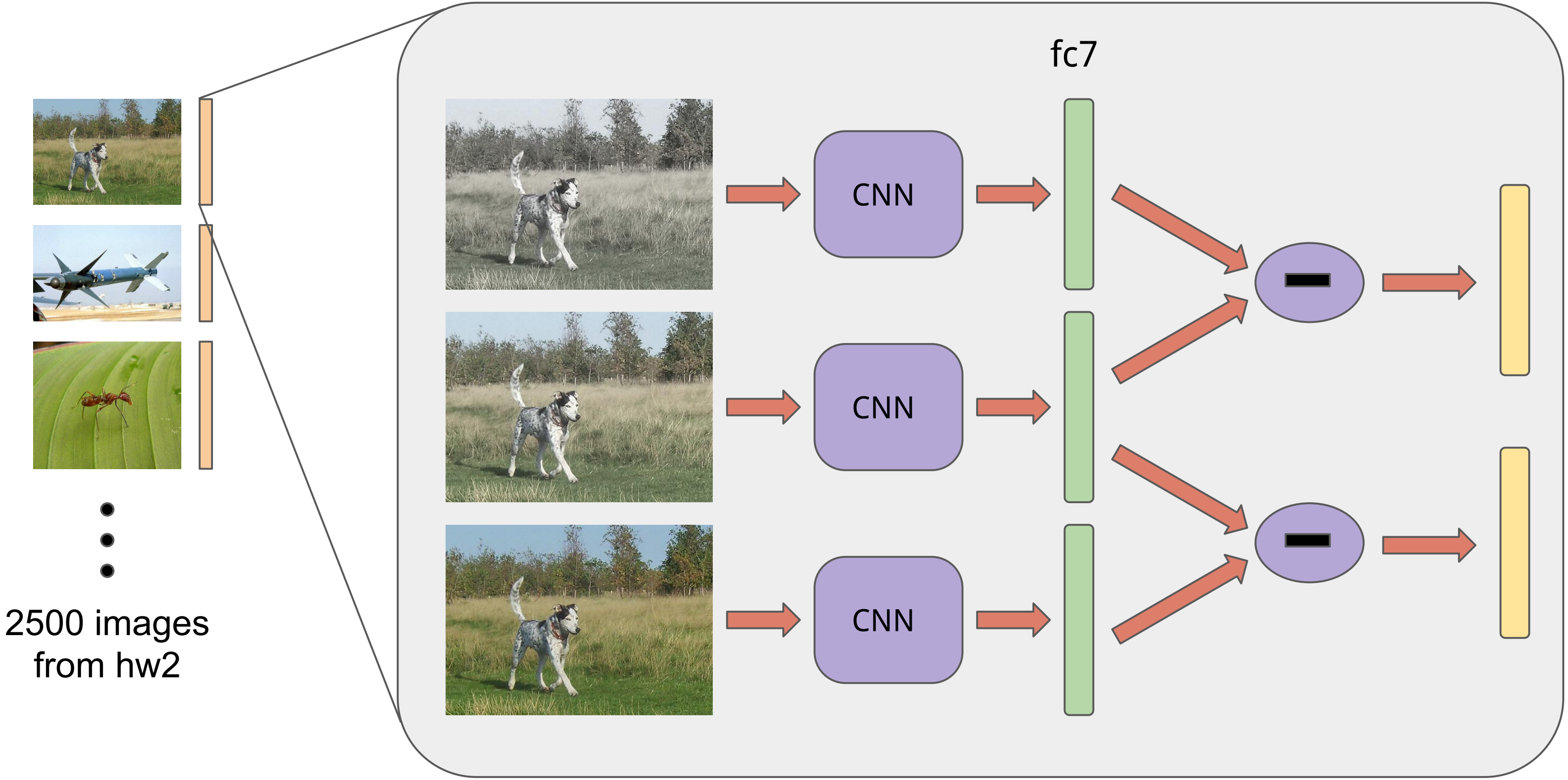

2500 images
from hw2

# Input variation - illumination

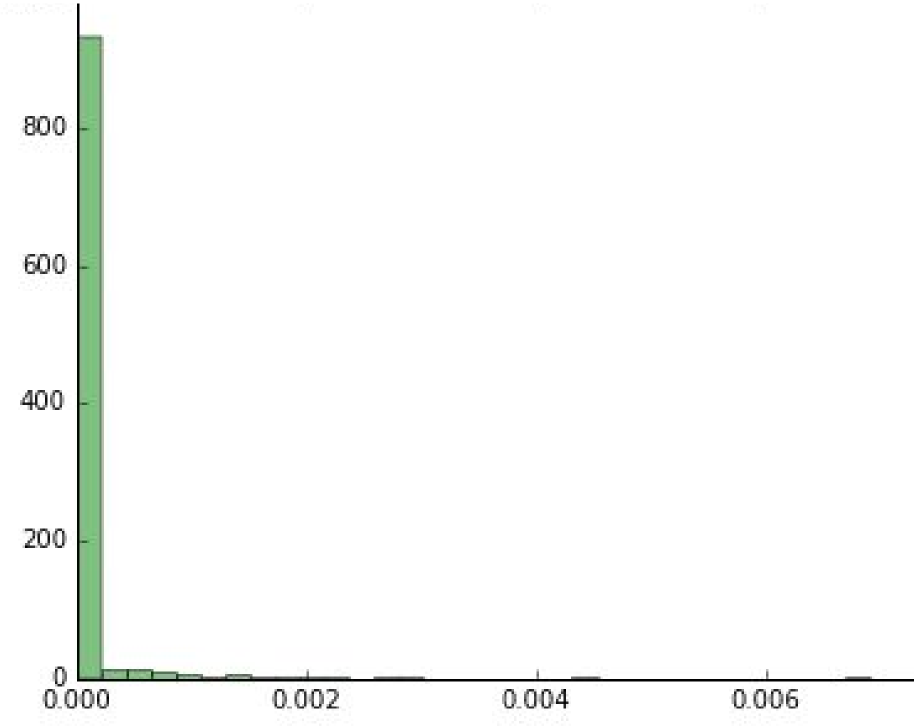


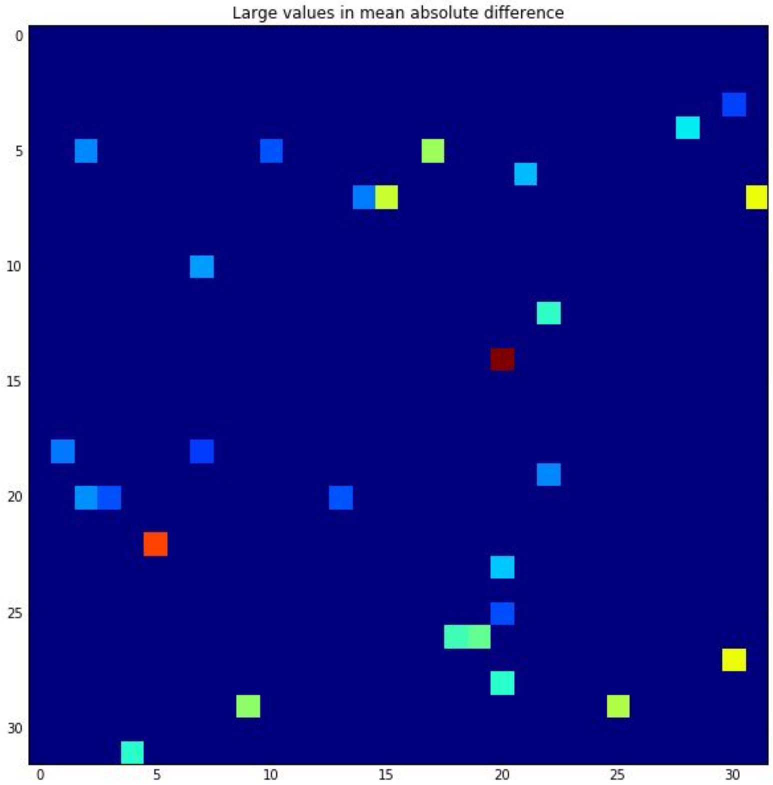



# Input variation - saturation



# Input variation - saturation



# Experiments - Outline

- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# Savings in labeling effort

- We want very good system even if it is expensive to collect labels
- If we finetune from the network in this paper, can we do away with less number of training examples?

| | **Performance** | **Comparison** | **Performance** |
|---|---|---|---|
| PASCAL VOC | 52% mAP | RCNN with AlexNet | 54.4% mAP |
| hw2 problem | 54.1% acc | Best non-finetuned model from hw2 | 52.8% acc |
| ImageNet - 10 | 4.9% acc | AlexNet - 10 | 0.15% acc |
| ImageNet - 100 | 15% acc | AlexNet - 14000 | 62.5% acc |

# Savings in labeling effort - discussion

- Unsupervised pretraining avoids overfitting
- 15% >> 0.1% random chance
- Tremendous in class variability in ImageNet. 100 images not sufficient to capture all of it
- PASCAL VOC results is for bounding boxes. ImageNet images can be the whole scene.
- PASCAL VOC has more than 100 images per class
- Should try with images per class

# Experiments - Outline

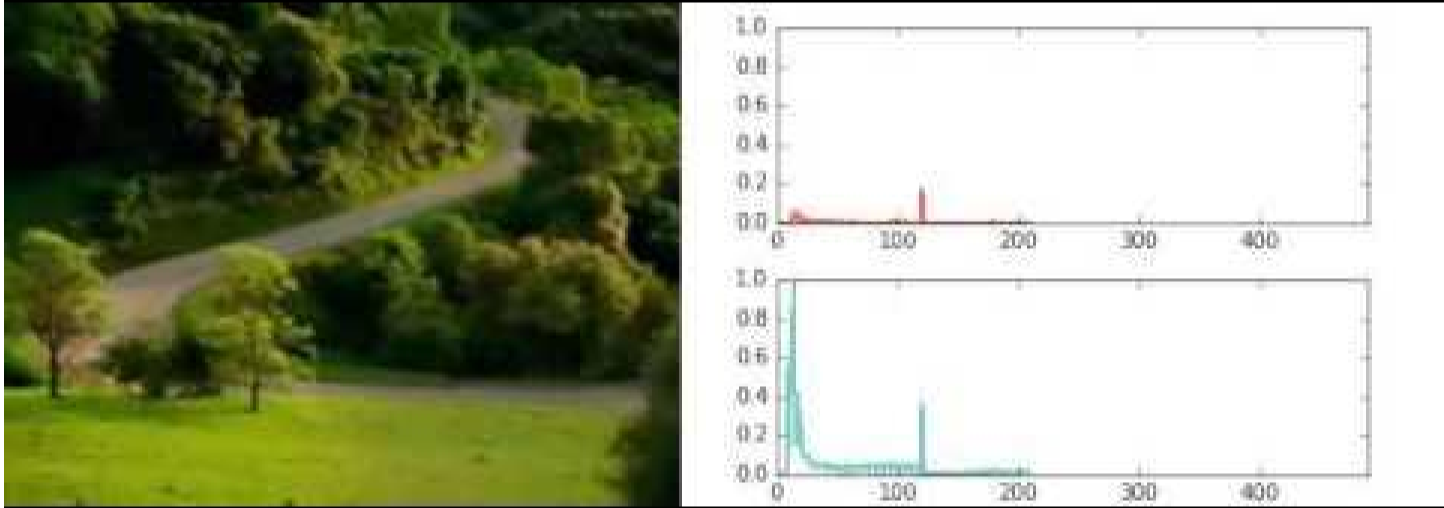
- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# Change point detection

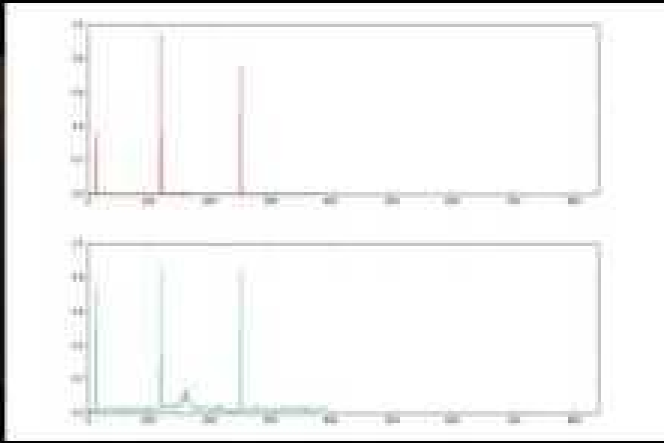
- Tracked patches from same video were used in paper
- Can create bias towards giving same representation to objects that appear together
- This experiment tests whether we can detect change points in the same video
- Very simple model : Magnitude of difference of embedding vectors of consecutive frames

Video 1

Video 1
Result



Video 2
Result



# Change point detection - discussion

As compared to embedding vector method, HoG baseline:

- gives larger changes when there is no visual change [start of car video]
- is more sensitive to occlusions [eg. white shirt entering]
- is more noisy even in stable sections of video

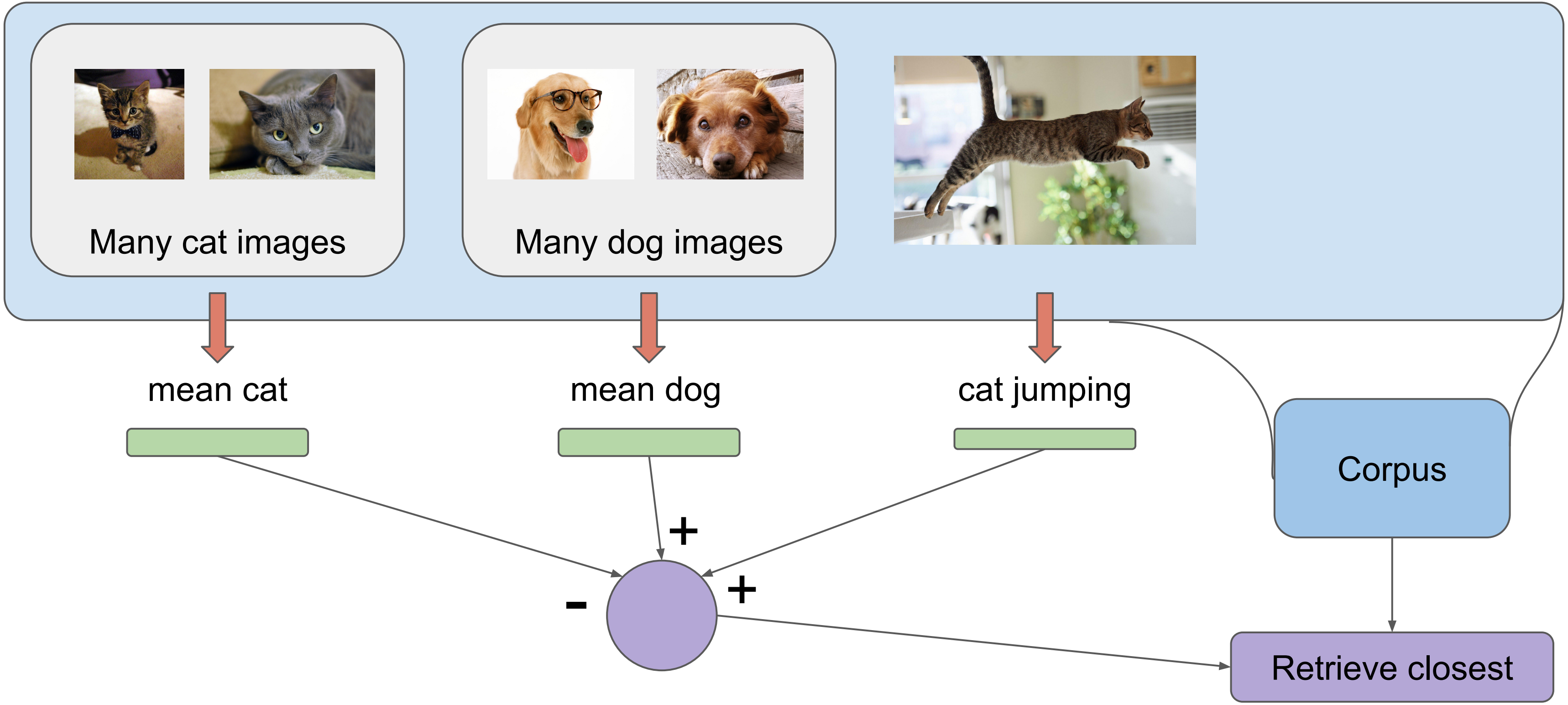
# Experiments - Outline

- tSNE visualization
- Effect of input variation
- Quantifying savings in labeling efforts
- Change point detection
- Relationship learning
- Discussion

# Relationship Learning

- Cosine similarity metric used during learning : similar to word2vec
- In word2vec: king - man + woman ≈ queen
  Do we have a similar thing here?
- Unlike word2vec, context is not explicitly provided but enters indirectly through temporal co-occurrence
- Idea : Use activity as context
  Example : cat_jumping - cat + dog ≈ dog_jumping?

# Relationship Learning : Small experiment




Images taken from Google Images

# Relationship Learning Results - top 3

- Should we be impressed?
  - No apparent similarity apart from similar action pose
  - The second image has very similar texture to first => honest mistake?
- Caveats
  - Single data point
  - Need a quantitative baseline

Images taken from Google Images

# Discussion

- This representation does not seem to capture activity very well.
  Possible solution : Learn embedding for video tubes instead of frames
- [Ramanathan et al] consider the whole image, while this one tracks patches across frames. Do we learn better representations with this?
- If this network is largely trained on moving objects, it can have little knowledge about the background or static scenes. This might affect its performance : tSNE plots seem to indicate otherwise
- Is most of the work in supervised part while finetuning?
  Best unsupervised was 44%, unsupervised learns good prior for finetuning
- Can we use audio to improve unsupervised learning?