Learning and Understanding Deep Visual Representations

Konda Reddy Mopuri and R. Venkatesh Babu

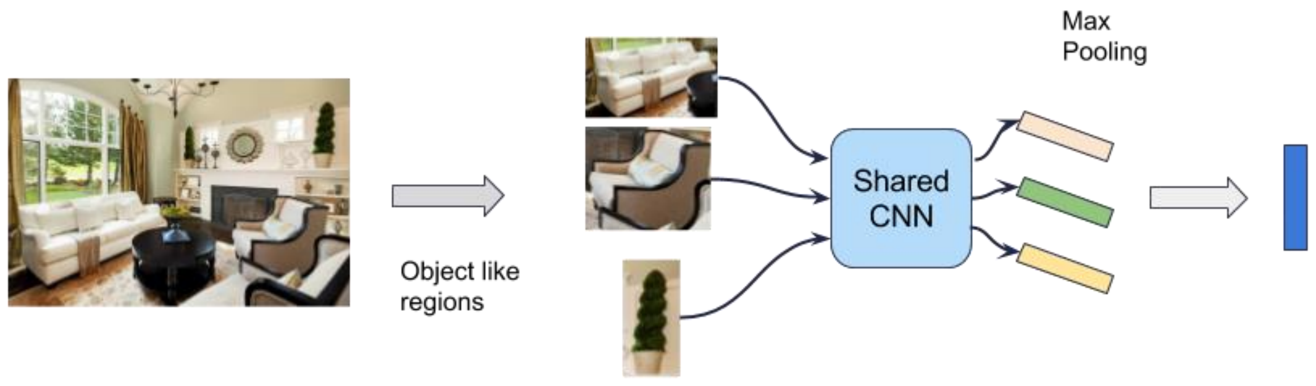
Video Analytics Lab, CDS, IISc

Learning deep visual representations from side and additional information

- Convolutional Neural Networks have resulted in unprecedented performances
- CNNs learn recognition from large scale datasets that offer category labels
- We exploit the useful "side and additional information" to enrich the representations with more semantics.
  - "Objectness"
  - Textual tags associated with images
  - Strong supervision offered by the captions.

Encoding "Objectness"

- Objects compose scenes → Detect and describe objects → scene summary



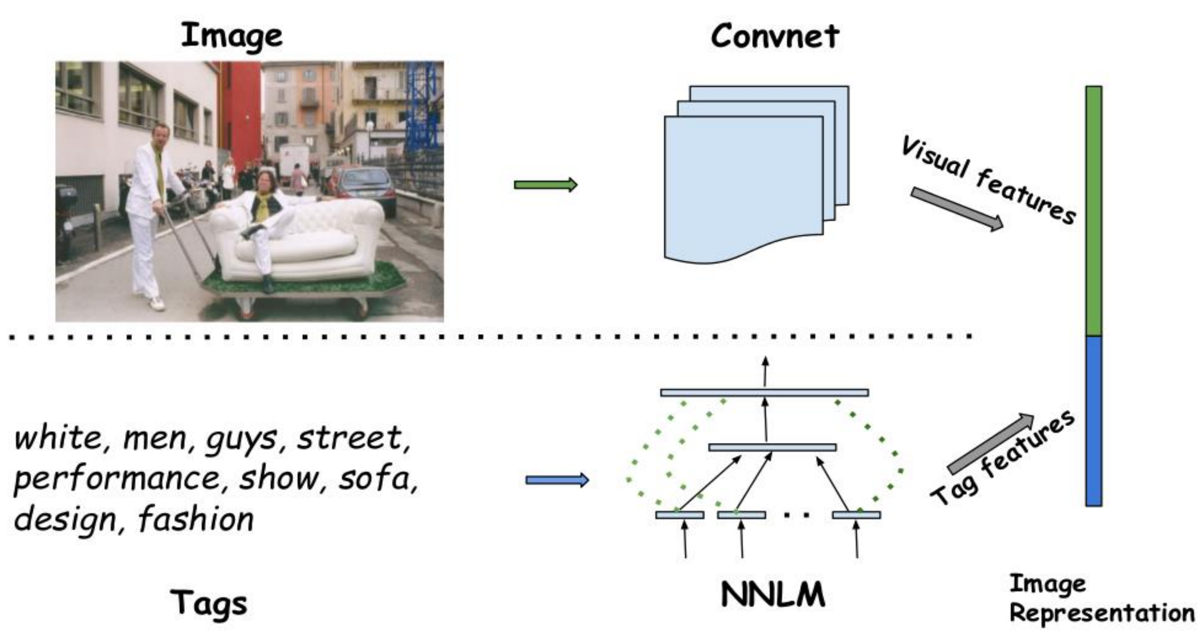
- CNNs' recognition is remarkable and show robustness to
  - Occlusion, pose, scale, intra class variation, etc.
- Same size as the CNN embeddings → dimensionality reduction

Table 1. Retrieval results on the *Holidays* dataset. Best performances in each column are shown in bold. (∇ *indicates result obtained with manual geometric alignment and retraining the CNN with similar database.*) Numbers indicate mAP (mean average precision).

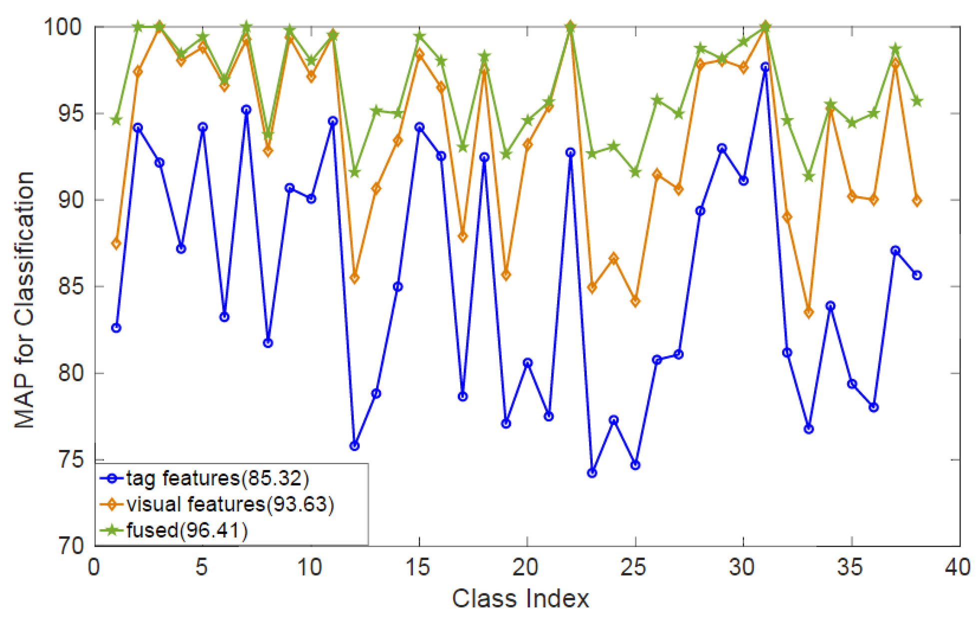
	Dimension									
METHOD	32	64	128	256	512	1024	2048	4096	8064	$\geq 10K$
VLAD	48.4	52.3	55.7	-	59.8	-	62.1	55.6		
Fisher Vector	48.6	52	56.5	-	61	-	62.6	59.5		
VLAD +adapt+ innorm	-	-	62.5	-	-	-	-	-	-	64.6
Fisher+color	-	-	-	-	-	-	-	77.4		
Multivoc-VLAD	-	-	61.4	-	-	-	-	-		
Triangulation Embedding	-	-	61.7	-	-	72.0	-	-	77.1	
Sparse-coded Features	-	-	0.727	-	-	-	-	-	-	76.7
Neural Codes	68.3	72.9	78.9^{∇}	74.9	74.9	-	-	*79.3^{∇}*		
MOP-CNN	-	-	-	-	-	-	80.2	78.9		
gVLAD	-	-	77.9	-	-	-	-	81.2		
Proposed	**73.96**	**80.67**	**85.09**	**87.77**	**88.46**	**86.58**	**85.94**	**85.94**		

Encoding Textual tags

- Images on web surrounded by rich text → multi-modal nature
- Encode (via language descriptors) and pool → representation from text
- Learn a classifier on top of the multi-modal representation

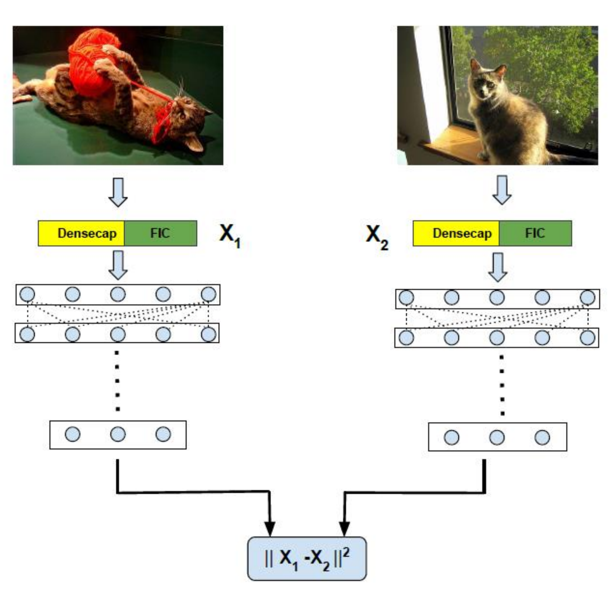


- Classification performance on MIRFlickr dataset
- 25K images from 39 concepts



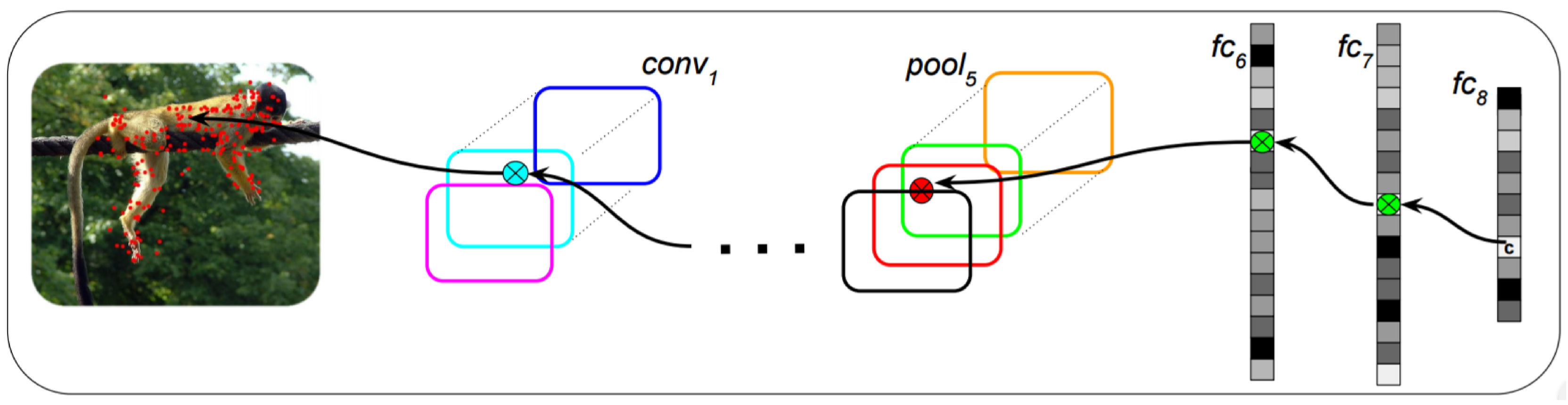
Stronger Supervision

- Transfer learning is common
- Current models are pre-trained on objects/scenes
- Applications like retrieval require scene summary
- Labels are not strong to summarize scene
- Explore transfer learning from caption generators (FIC) and region descriptors (DenseCap)
- Impose pairwise constraints

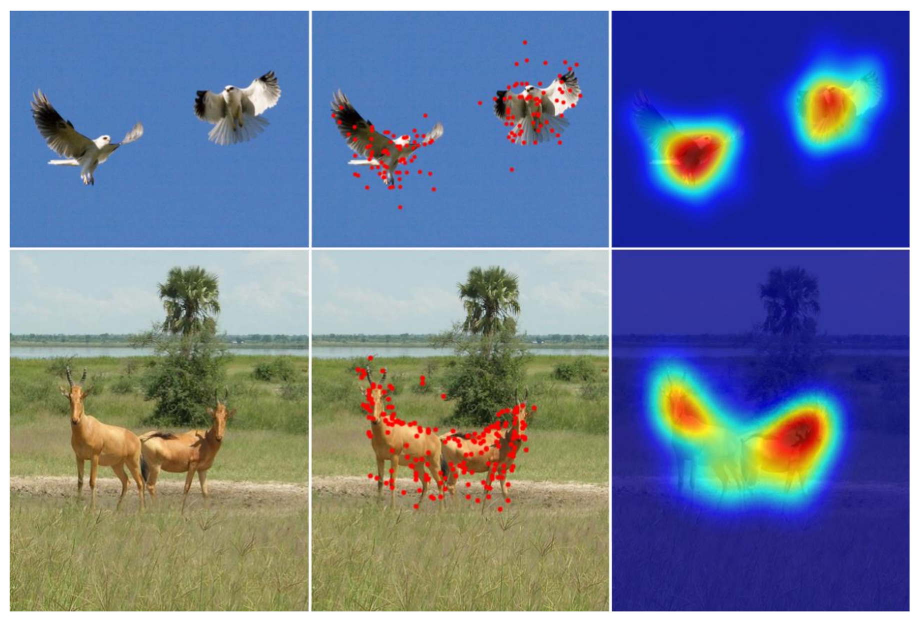


Understanding deep visual representations

- In spite of impressive performance, CNNs offer limited transparency
- Therefore treated as black-boxes
- One way to understand → determine the important image locations that guide the CNN's prediction
- We exploit the feature dependencies across the layers to locate the evidence

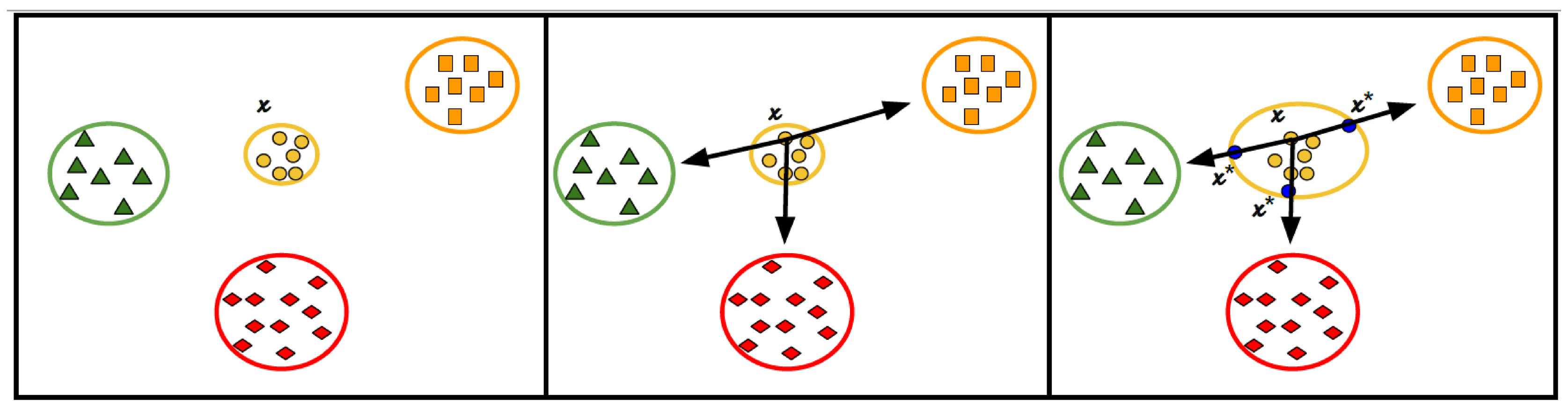


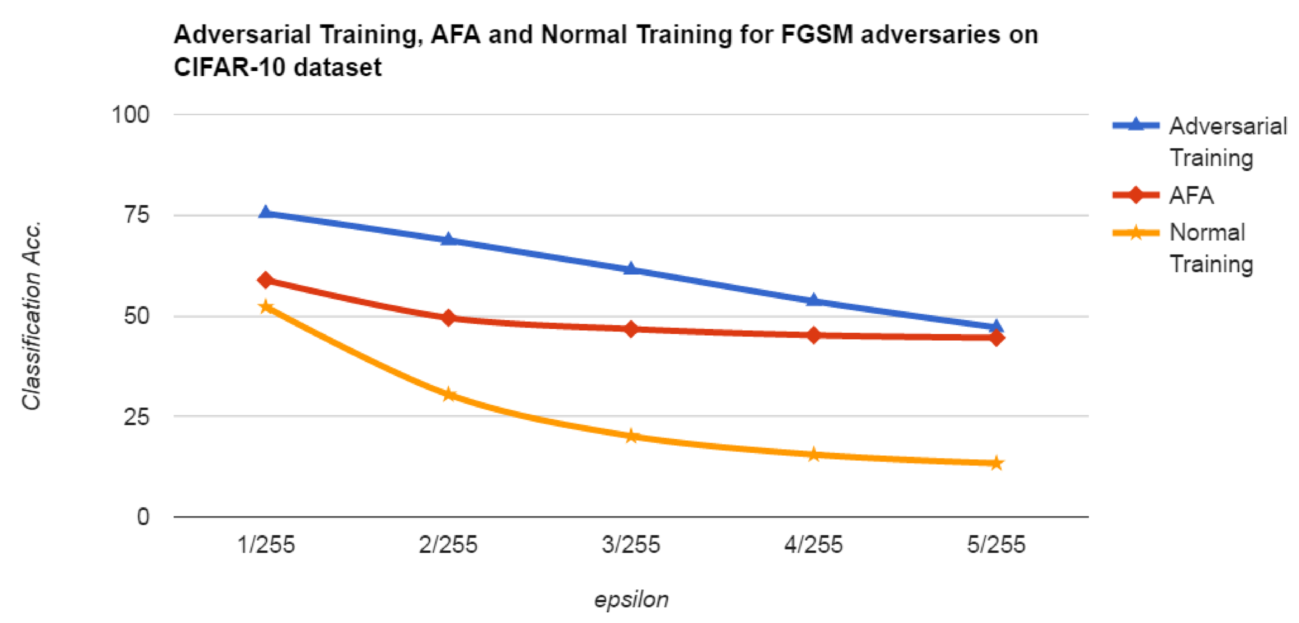
- Provides visual explanations
- Perform
  - Weakly supervised localization
  - Saliency
  - Caption grounding



Adversarial Feature Augmentation (AFA) for learning robust models

- Adversarial samples fool ML systems – CNNs are no exception
- Solution: Adversarial training is required
  - Train with adversarial + normal samples
  - Highly inefficient
- We propose a feature level augmentation technique to handle ANY adversarial attack
- Augment the embeddings (data) into adversarial directions
- Include them in the training with the original labels





Conclusions

- Deep learned image representations can be made more discriminative and useful via augmenting with task specific side information and additional semantic information
- Despite their excellent performance, CNNs leave various design aspects opaque. Visualizing the predictions can help develop more useful insights into the design and training of these complex ML systems.
- Adversarial feature augmentation can help CNNs learn smoother mappings and make them robust to multiple adversaries.

Learning and Understanding Deep Visual Representations

Konda Reddy Mopuri and R. Venkatesh Babu

Video Analytics Lab, CDS, IISc

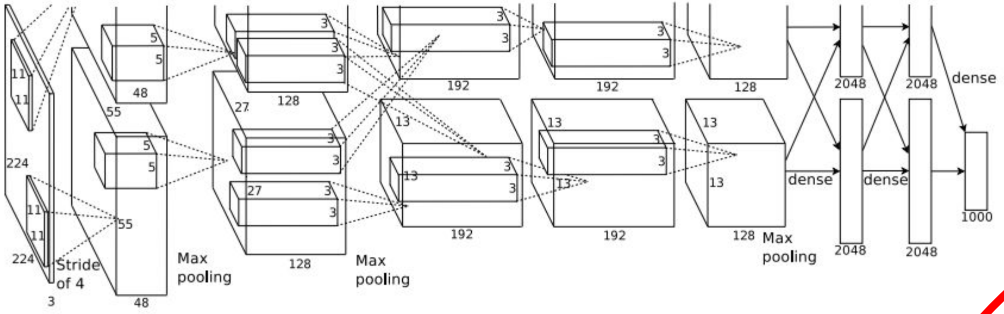
What we will discuss

- Learning visual representations using Convolutional Neural Networks
  - Side and additional information
    - objectness, textual tags, etc.
  - How can we encode it to augment the semantics ?
- Understanding the representations
  - Visual explanations for predictions
  - Adversarial augmentation

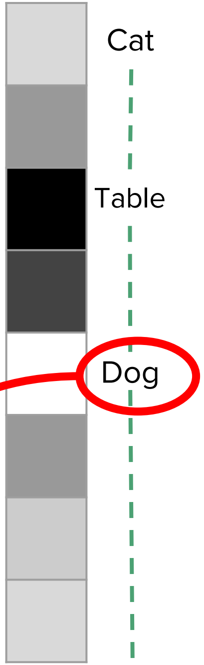
Representation learning using CNNs

DATA

Image



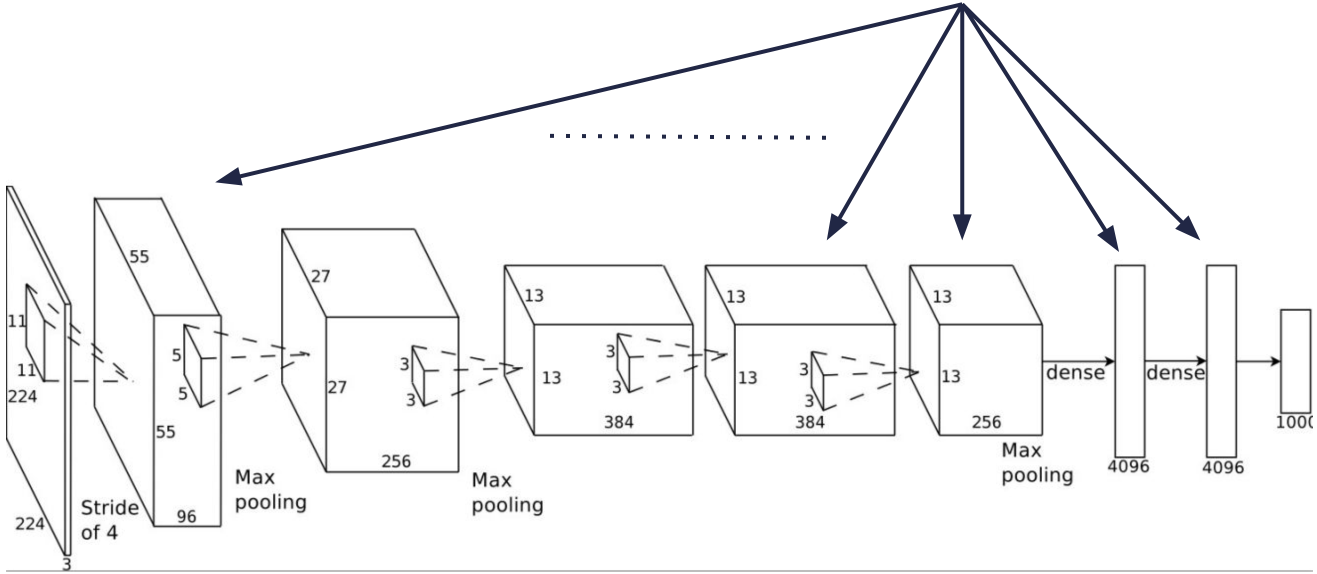
Lot of annotated data



Class Probabilities

The image representations

Hierarchy of features



The side and additional info.

- **Objectness**
- Textual tags
- Strong supervision

The practical issues

- Typical images contain multiple (unseen) objects (scales), scenes

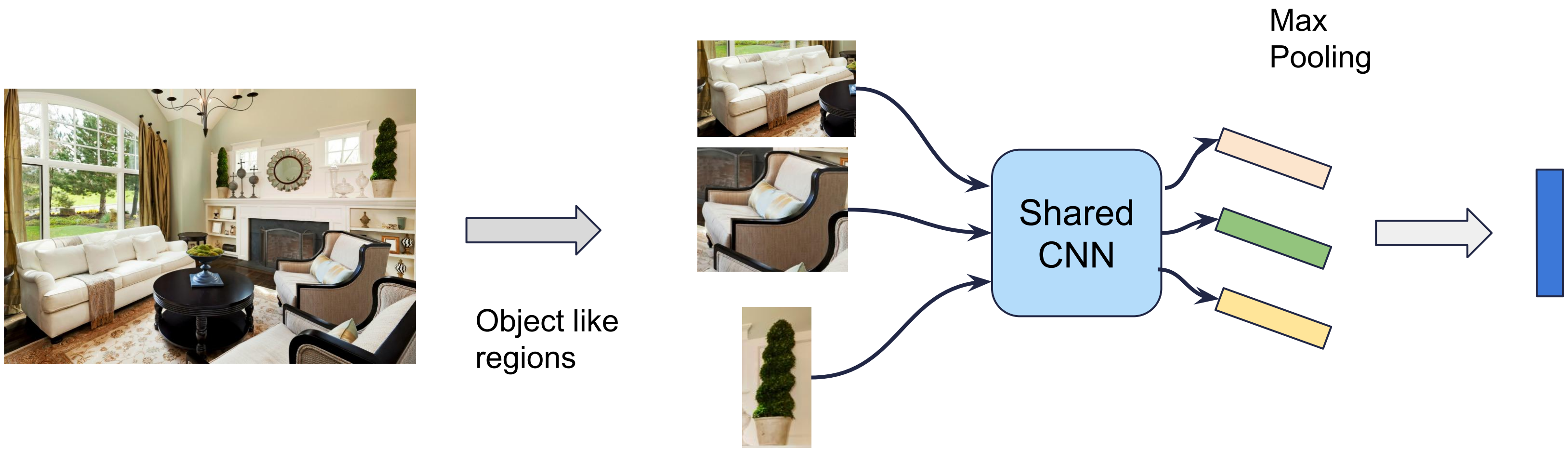
$\neq$

Typical real world images

Dataset train images

Objectness prior

- Objects compose scenes
- Detect and describe objects ⟶ scene summary



Object level deep feature pooling for compact Image representation, Konda Reddy Mopuri et al. CVPRW 2015

Results

Table 1. Retrieval results on the *Holidays* dataset. Best performances in each column are shown in bold. (∇ *indicates result obtained with manual geometric alignment and retraining the CNN with similar database*.)

	Dimension									
METHOD	32	64	128	256	512	1024	2048	4096	8064	$\geq 10K$
VLAD [16]	48.4	52.3	55.7	-	59.8	-	62.1	55.6		
Fisher Vector[27]	48.6	52	56.5	-	61	-	62.6	59.5		
VLAD +adapt+ innorm [1]	-	-	62.5	-	-	-	-	-	-	64.6
Fisher+color [13]	-	-	-	-	-	-	-	77.4		
Multivoc-VLAD [14]	-	-	61.4	-	-	-	-	-		
Triangulation Embedding [17]	-	-	61.7	-	-	72.0	-	-	77.1	
Sparse-coded Features [9]	-	-	0.727	-	-	-	-	-	-	76.7
Neural Codes [2]	68.3	72.9	78.9^{∇}	74.9	74.9	-	-	*79.3*$^{\nabla}$		
MOP-CNN [12]	-	-	-	-	-	-	80.2	78.9		
gVLAD [33]	-	-	77.9	-	-	-	-	81.2		
Proposed	**73.96**	**80.67**	**85.09**	**87.77**	**88.46**	**86.58**	**85.94**	**85.94**		

Object level deep feature pooling, Konda Reddy Mopuri et al. CVPRW 2015

The side and additional info.

- Objectness
- **Textual tags**
- Strong supervision

Images on web

- Surrounded by rich text → multi-modal nature

"five large white wind turbines are standing on a dark green slope connected by brown dirt roads"

Image

white, men, guys, street, performance, show, sofa, design, fashion

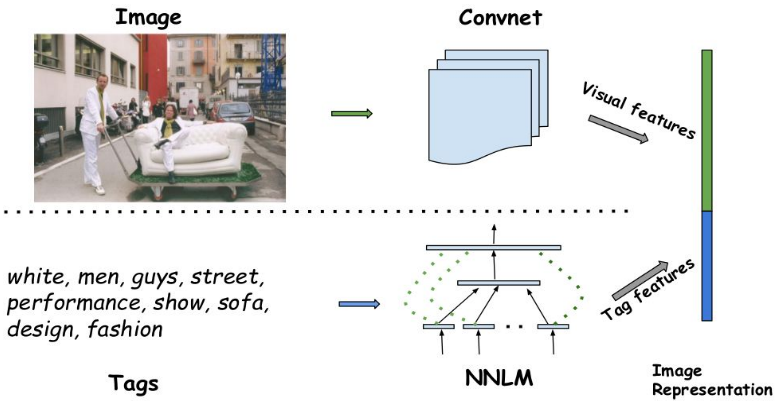
Tags

Extract additional semantics

- Traditional methods - BoW encoding
  - Inefficient, Not semantic
- Cross modal/Joint learning - CCA, CorrNets, etc.
  - Scaling issues
- Neural nets based language models
  - Semantic and preserve regularities
  - Ex: Word2vec, glove, thought vectors, etc.

Late fusion

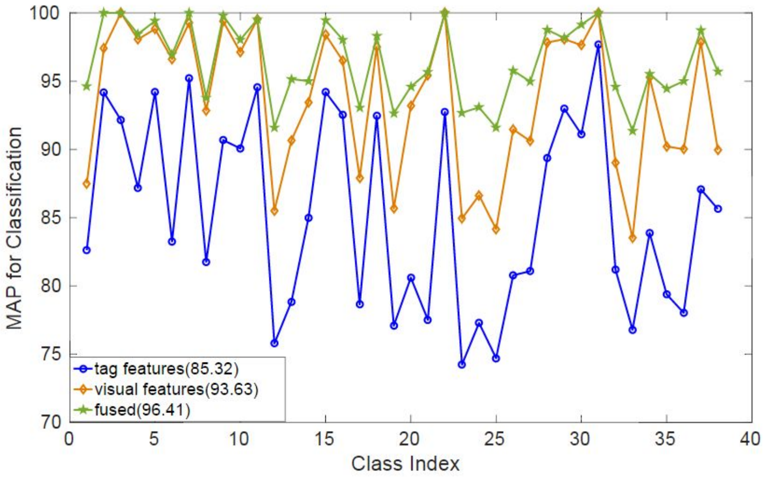
- Preprocess
  - Noise and stop words removal, lemmatizing
- Encode and pool → representation from text
- Learn a classifier on top of the multi-modal representation



Konda Reddy M et al., "Towards Semantic Visual Representation: Augmenting Image Representation with Natural Language Descriptors" (ICVGIP 2016)

Drawbacks

- Noisy tags
- Visual similarity vs learned semantics
  - Ex: Cat and dog might have similar embeddings from text
- Minor improvements



The side and additional info.

- Objectness
- Textual tags
- **Strong supervision**

Labels are weak

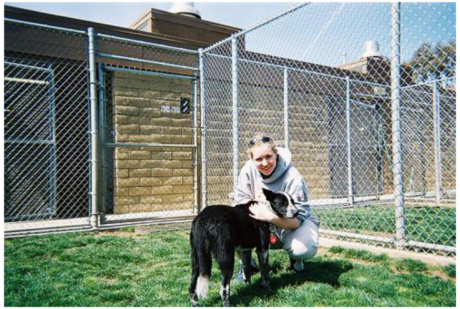
- ❖ Current models are pretrained on objects/scenes
- ❖ Applications like retrieval require whole scene summary
- ❖ Labels ⟶ Not strong enough to summarize scene

dog

fish

cup

Dataset train images



Sample query image

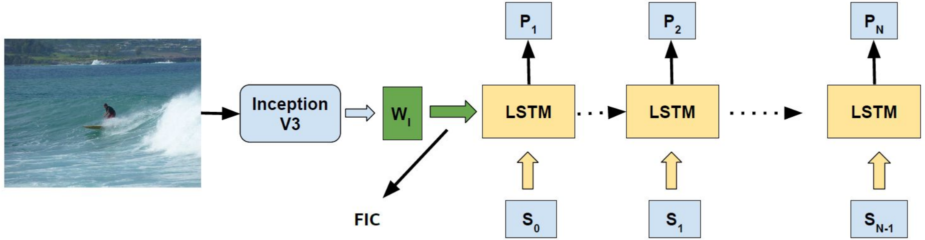
We need strong supervision

A man holding a child while standing at the fence of an elephant zoo enclosure.

The horse and puppy are separated by the mesh fence.

Caption generators (FIC) & Dense region descriptors (DenseCap)

A man riding a wave on top of a surfboard



a man and a dog. a green grassy field. a metal fence. a black and white dog. a brick wall. metal fence behind the fence. man wearing a white shirt. green grass on the ground. man wearing a white shirt. a black and white dog.

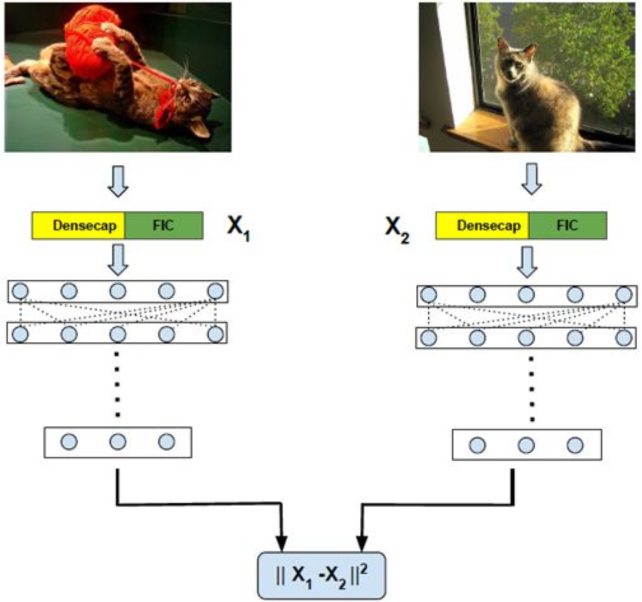
Transfer learning from caption generators

- Exploit the features learned by the systems with strong supervision
- Transfer their learning for our target applications
- Perform learning on top with task specific constraints

Transfer learning with task specific constraints

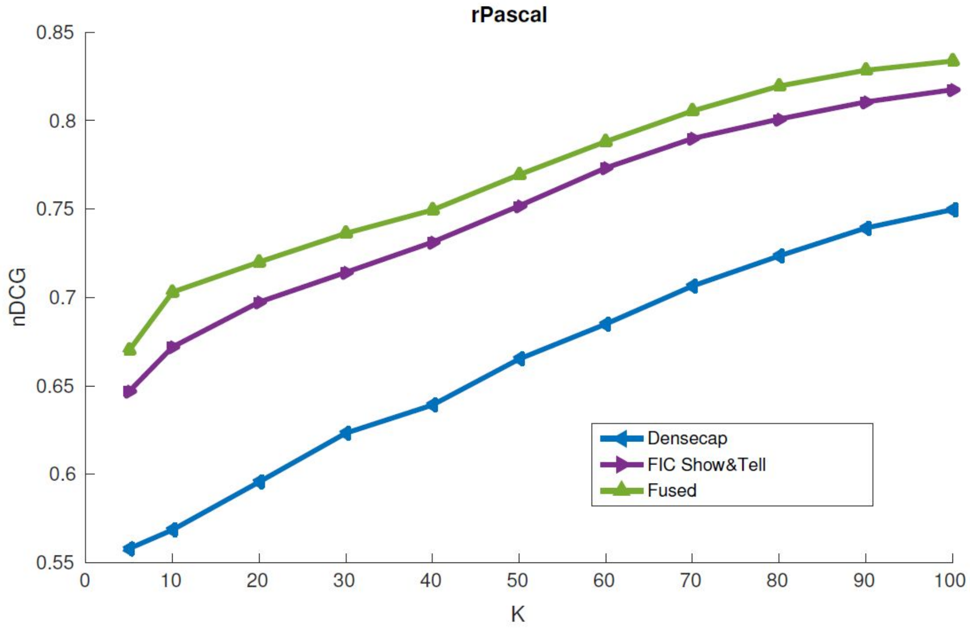
$$E = \frac{1}{2N} \sum_{n=1}^{N} (y^2)\, d + 1\,(y=0) \max\left(\nabla - d^2, 0\right)$$

Modified siamese loss



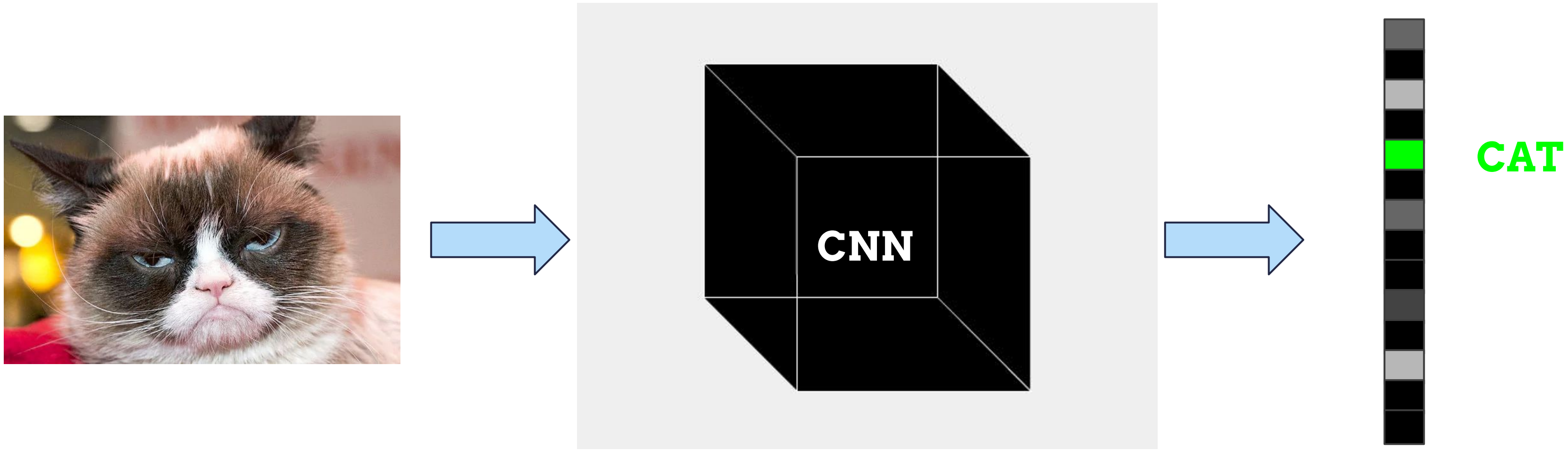
Konda Reddy M et al., "Deep Image Representations using Caption generators" (ICME 2017)

Retrieval results



Understanding the representations

CNNs are black boxes ?



Interpretability matters

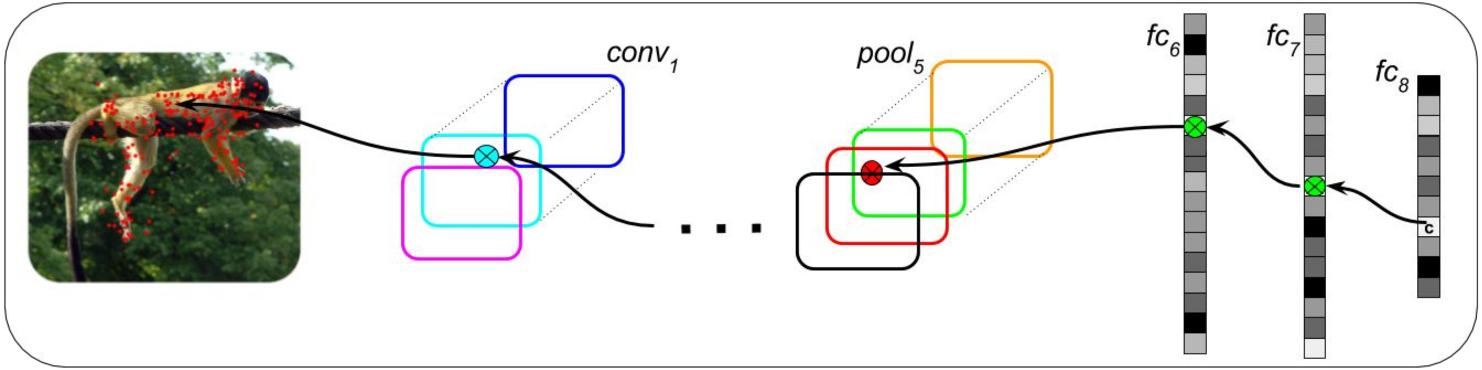
- Lack of decomposability
- No transparency
  - when they fail → no warning, no explanation
- Suffer from the trade-off b/w "Accuracy" and "Interpretability"

Additional information

- Reason an inference
- Visual explanations

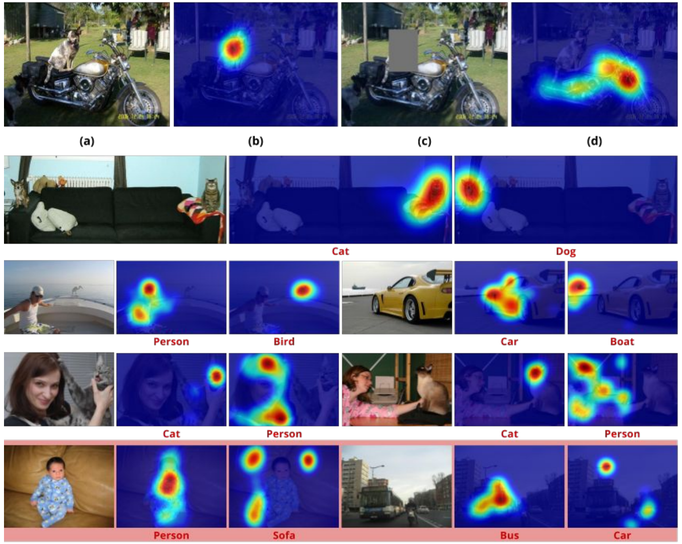
CNN Fixations

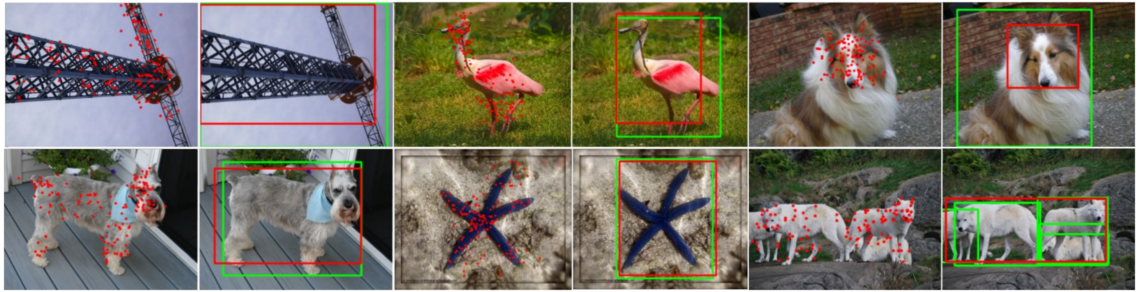
- Exploit the feature dependencies to locate the evidence
- Iteratively backtrack onto the image from the predicted label

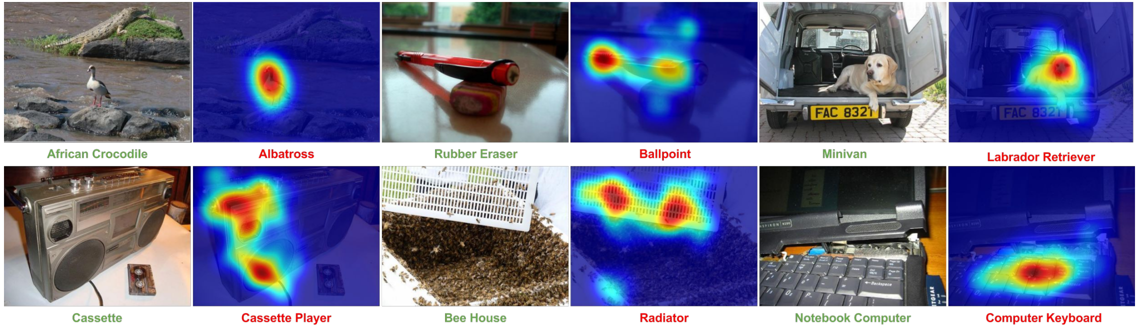


(a)

CNN-Fixations are useful







Adversarial Images

Images that fool CNNs

Shetland sheepdog

Paintbrush

Why do they exist ?

- Multiple hypotheses
  - Linearity, low probability pockets in the feature space, sparse training data, etc.
- Multiple methods to generate them

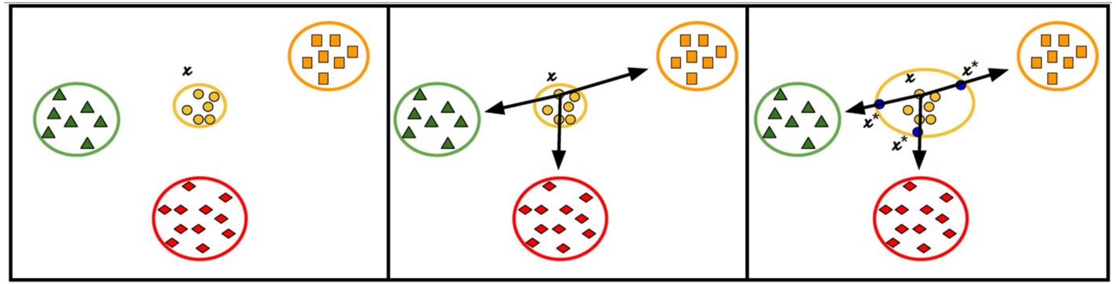
Current ways of handling it

- Adversarial training
- Each iteration → Train with Normal + Adversarial images
- Highly inefficient
  - Compute adversarial images at each iteration

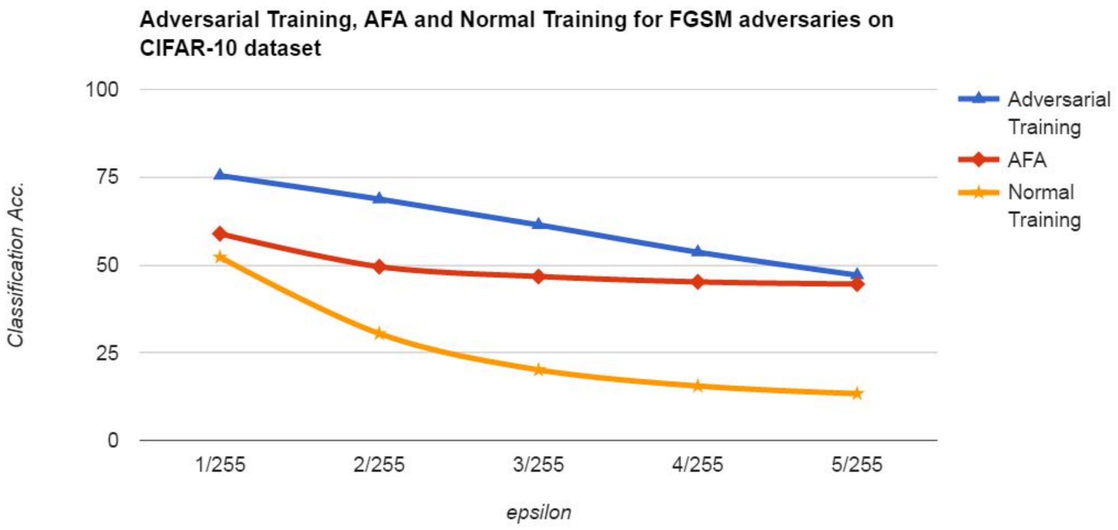
Feature level augmentation

- Expand features in the embedding space
- Choose random adversarial directions
- Replace normal data with augmented data
  - No extra computations
  - No knowledge about the nature of attack

$$X^{aug} = X + \eta * (X_p - X)$$



The tradeoff



Conclusions

- ❖ Deep learned image representations can be made more discriminative and useful
  - ➢ Augment with task specific side information and additional semantic information
- ❖ Visualizing the predictions can help to develop more useful insights into the design and training of these complex ML systems.
- ❖ "Adversarial images" is an intriguing aspect of ML systems that demands rigorous study.

Thank you.