DeCAF: a Deep Convolutional Activation Feature for Generic Visual Recognition

J Donahue*, Y Jia*, O Vinyals, J Hoffman, N Zhang, E Tzeng, T Darrell.

BVLC

Berkeley
UNIVERSITY OF CALIFORNIA

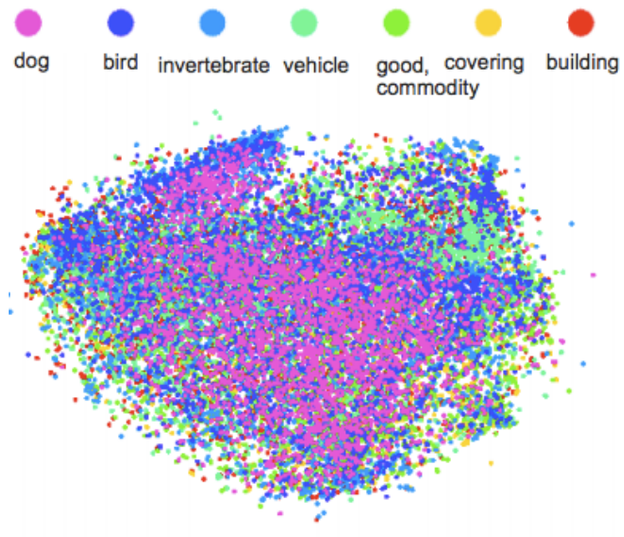
Features Representation's challenge

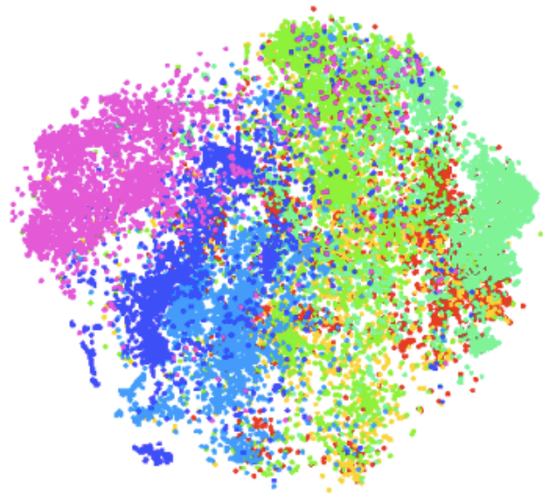
PROBLEM

- performance with conventional visual representations (*flat feature representations*) has been impressive but has likely plateaued in recent years

SOLUTION

- discover effective representations that capture salient semantics for a given task
- deep architectures should be able to do this





A little bit of History

- Deep CNN has a long history in computer vision
  - supervised back-propagation networks to perform digit recognition *[LeCun et al., 1989]*

- Recently CNN have achieved competition-winning numbers on large benchmark dataset
  - convolutional network proposed by Krizhevsky (2012)
  - dataset consisting of more than one million images (ImageNet) *[Berg et al., 2012]*

- Learning from related tasks has also a long history in machine learning *[Caruana, 1997 - Argyriou et al., 2006]*

- In computer Vision forming a representation based on sets of trained classifiers on related tasks has recently show to be effective *[Torresani et al., 2010 - Li et al., 2010]*

PROBLEM

- Transfer learning using deep representation bad in unsupervised setting
  - limited with relatively small datasets (CIFAR and MNIST)
  - modest success on larger datasets *[Le et al., 2012]*

Why Deep Models

- deep or layered compositional architectures should be able to capture salient aspects of a given domain [*Krizhevsky NIPS 2012*][*Singh ECCV 2012*]

- perform better than traditional hand-engineered representations in many domains
  - especially where good features has not already been engineered [*Le CVPR 2011*]

- recently applied to large-scale visual recognition tasks
  - recently outperformed all known methods on a large scale recognition challenge
  - performs extremely well in domains with large amounts of training data

HOWEVER

- with limited training data, fully-supervised deep architectures generally overfit
- many conventional visual recognition challenges have tasks with few training examples

Idea

- investigate a deep architecture
  - representations are learned on a set of related problems
  - applied to new tasks which have too few training examples

- model considered as a deep architecture for *transfer learning*
  - based on a supervised pre-training phase
  - new visual features "*DeCAF*" defined by convolutional network weights

WHY

- empirical validation
  - that generic visual feature based on a CNN weights trained on ImageNet outperforms conventional visual representations

WITH

- Caltech-101 (Object recognition dataset *[Fei-Fei et al., 2004]*)
- Office (Domain adaptation dataset *[Saenko et al., 2010]*)
- Caltech-UCSD (Birds fine-grained recognition dataset *[Welinder et al., 2010]*)
- SUN-397 (Scene recognition dataset *[Xiao et al., 2010]*)

Approach

- Train a Deep convolutional model in a fully supervised setting using Krizhevsky method
  - state-of-the-art method
  - large scale dataset for training (ImageNet)
- Extract various features from the network
- Evaluate the efficacy of these features on generic vision tasks

TWO IMPORTANT QUESTIONS

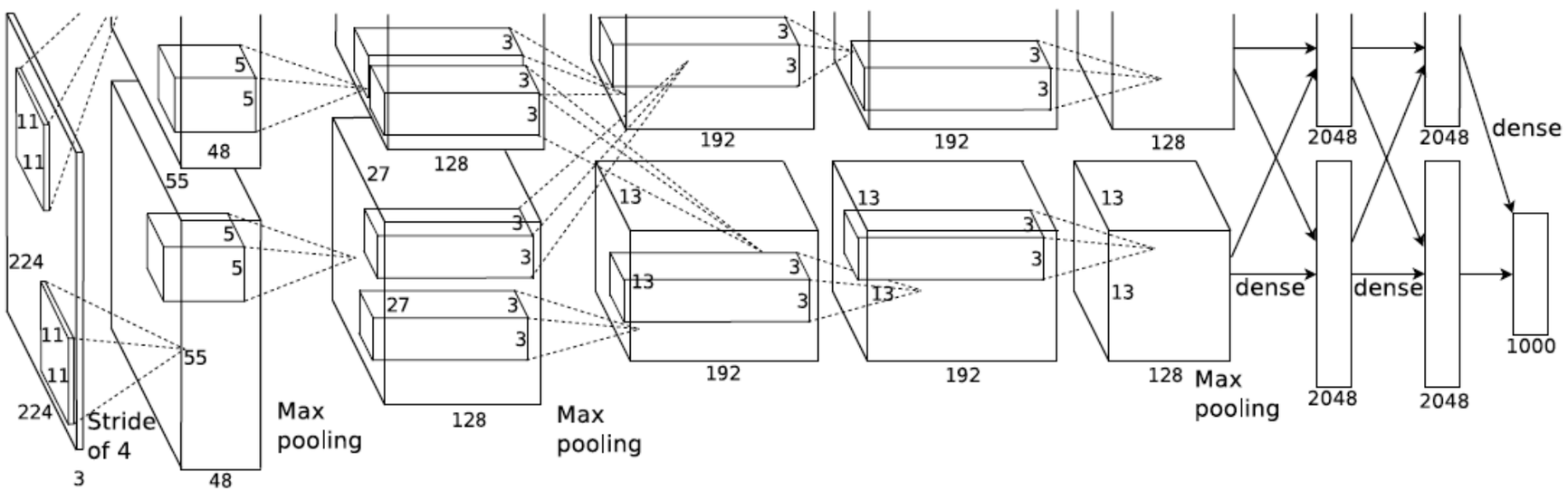
- Do features extracted from the CNN generalize the other datasets ?
- How does performance vary with network depth ?

FEEDBACK

- qualitatively and quantitatively via visualizations of semantic clusters
- experimental comparison to current baselines

Adopted Network

- Deep CNN architecture proposed by Krizhevsky et al. (2012)
  - values propagated through 5 convolutional layers (with pooling and ReLU)
  - 3 fully-connected layers to determinate final neuron activities
  - won ImageNet Large Scale Visual recognition Challenge 2012 *[Berg et al., 2012]*
  - top-1 validation error rate of 40.7%



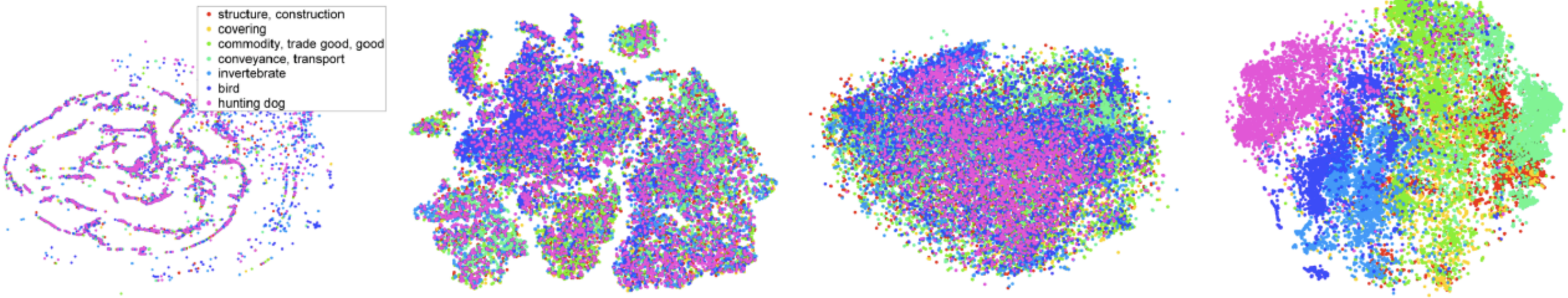
- follow architecture and training protocol with two differences
  - input 256 x 256 images rather than 224 x 224 images
  - no data augmentation trick (eg. adding random multiples of the p.c of the RGB)

Qualitatively and Quantitatively Feedback 1/2

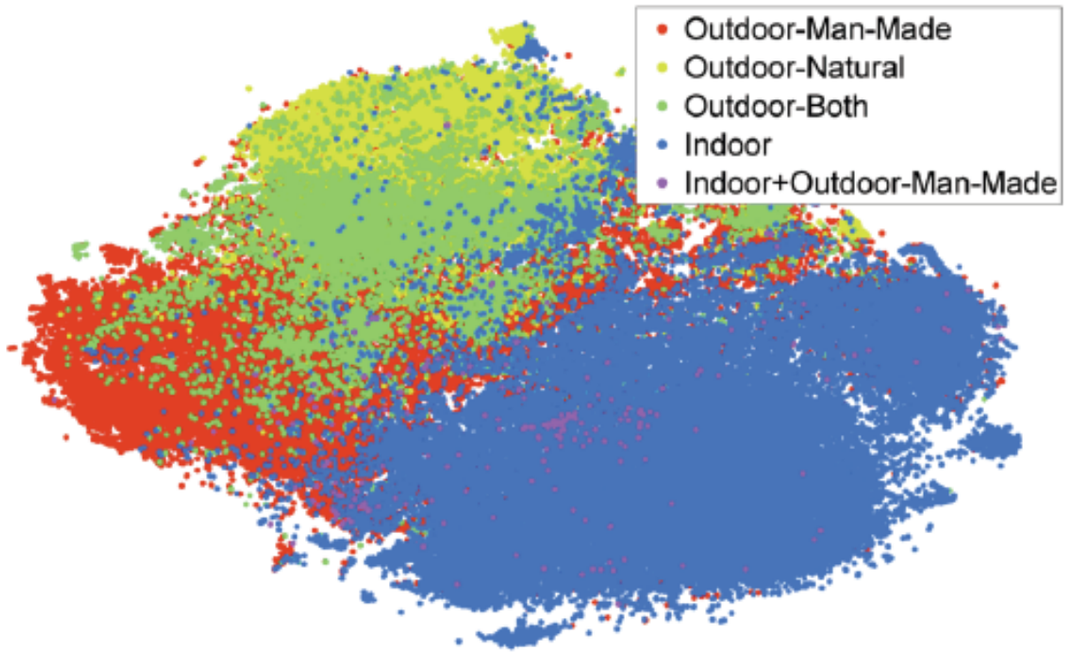
- To gain insight into the semantic capacity of DeCAF features
- Comparison with GIST features *[Oliva & Torralba, 2001]* and LLC features *[Wang at al., 2010]*
- Use of t-SNE algorithm *[van der Maaten & Hilton, 2008]*
  - find 2-dimensional embedding of the hight-dimensional feature space
  - plot as a points colored depending on their semantic category
- Use of ILSVRC-2012 validation set to avoid overfitting (150,000 photographs, collected from flickr and other search engines)
- Use of SUN-397 dataset to evaluate how dataset bias affects results

Qualitatively and Quantitatively Feedback 2/2

- Take the activations of *n* hidden layer of the CNN as a feature $DeCAF_n$



(a) LLC (b) GIST (c) $DeCAF_1$ (d) $DeCAF_6$



Experimental Comparison Feedback

- Experimental results evaluating DeCAF on multiple standard computer vision benchmarks

- Not evaluation of features from any earlier layers in the CNN
  - do not contain rich semantic representation

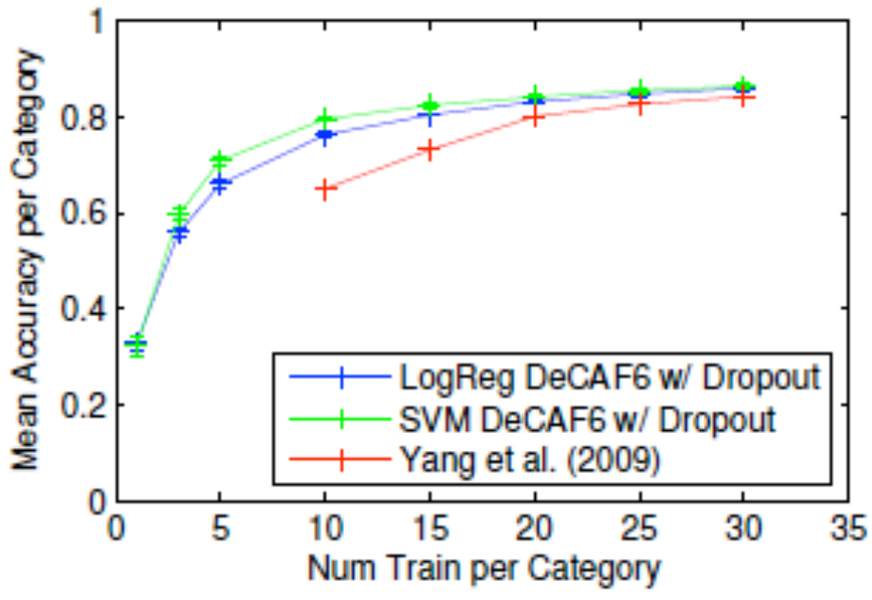
- Results on multiple datasets to evaluate the strength of DeCAF for
  - basic object recognition (Caltech-101)
  - domain adaptation (Office)
  - fine-grained recognition (Caltech-UCSD)
  - scene recognition (SUN-397)

- Together represent much of the contemporary visual recognition spectrum

Object Recognition

- Evaluation also of a regularization technique called "*dropout*" *[Hilton et al., 2012]*
- Classifier trained on random set of 30 samples per class and tested on the rest
- Results compared with current state-of-the-art on this benchmark *[Yang et al. 2009]*
  - combination of 5 traditional hand-engineered image features
- Compared also with the two-layers convolutional network of Jarret et al (2009)
  - to demonstrate the importance of the depth of the network used for this features

	$DeCAF_5$	$DeCAF_6$	$DeCAF_7$
LogReg	63.29 ± 6.6	84.30 ± 1.6	84.87 ± 0.6
LogReg with Dropout	-	86.08 ± 0.8	85.68 ± 0.6
SVM	77.12 ± 1.1	84.77 ± 1.2	83.24 ± 1.2
SVM with Dropout	-	$\mathbf{86.91 \pm 0.7}$	85.51 ± 0.9
Yang et al. (2009)		84.3	
Jarrett et al. (2009)		65.5	



Domain Adaptation 1/2

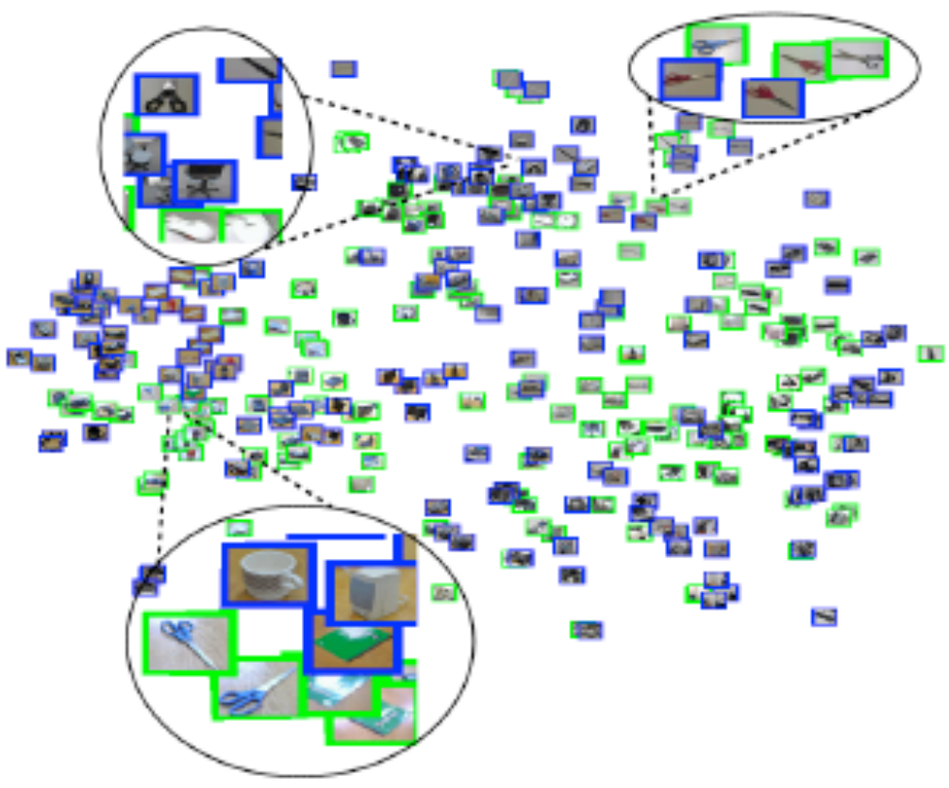
- Particular dataset used with three domains
  - *Amazon*: images taken from amazon.com
  - *Webcam* and *Dslr*: images taken in office environment using a webcam or SLR camera

	Amazon → Webcam			Dslr → Webcam		
	SURF	$DeCAF_6$	$DeCAF_7$	SURF	$DeCAF_6$	$DeCAF_7$
Logistic Reg. (S)	9.63 ± 1.4	48.58 ± 1.3	53.56 ± 1.5	24.22 ± 1.8	88.77 ± 1.2	87.38 ± 2.2
SVM (S)	11.05 ± 2.3	52.22 ± 1.7	53.90 ± 2.2	38.80 ± 0.7	91.48 ± 1.5	89.15 ± 1.7
Logistic Reg. (T)	24.33 ± 2.1	72.56 ± 2.1	74.19 ± 2.8	24.33 ± 2.1	72.56 ± 2.1	74.19 ± 2.8
SVM (T)	51.05 ± 2.0	78.26 ± 2.6	78.72 ± 2.3	51.05 ± 2.0	78.26 ± 2.6	78.72 ± 2.3
Logistic Reg. (ST)	19.89 ± 1.7	75.30 ± 2.0	76.32 ± 2.0	36.55 ± 2.2	92.88 ± 0.6	91.91 ± 2.0
SVM (ST)	23.19 ± 3.5	80.66 ± 2.3	79.12 ± 2.1	46.32 ± 1.1	$\mathbf{94.79 \pm 1.2}$	92.96 ± 2.0
Daume III (2007)	40.26 ± 1.1	$\mathbf{82.14 \pm 1.9}$	81.65 ± 2.4	55.07 ± 3.0	91.25 ± 1.1	89.52 ± 2.2
Hoffman et al. (2013)	37.66 ± 2.2	80.06 ± 2.7	80.37 ± 2.0	53.65 ± 3.3	93.25 ± 1.5	91.45 ± 1.5
Gong et al. (2012)	39.80 ± 2.3	75.21 ± 1.2	77.55 ± 1.9	39.12 ± 1.3	88.40 ± 1.0	88.66 ± 1.9
Chopra et al. (2013)		58.85			78.21	

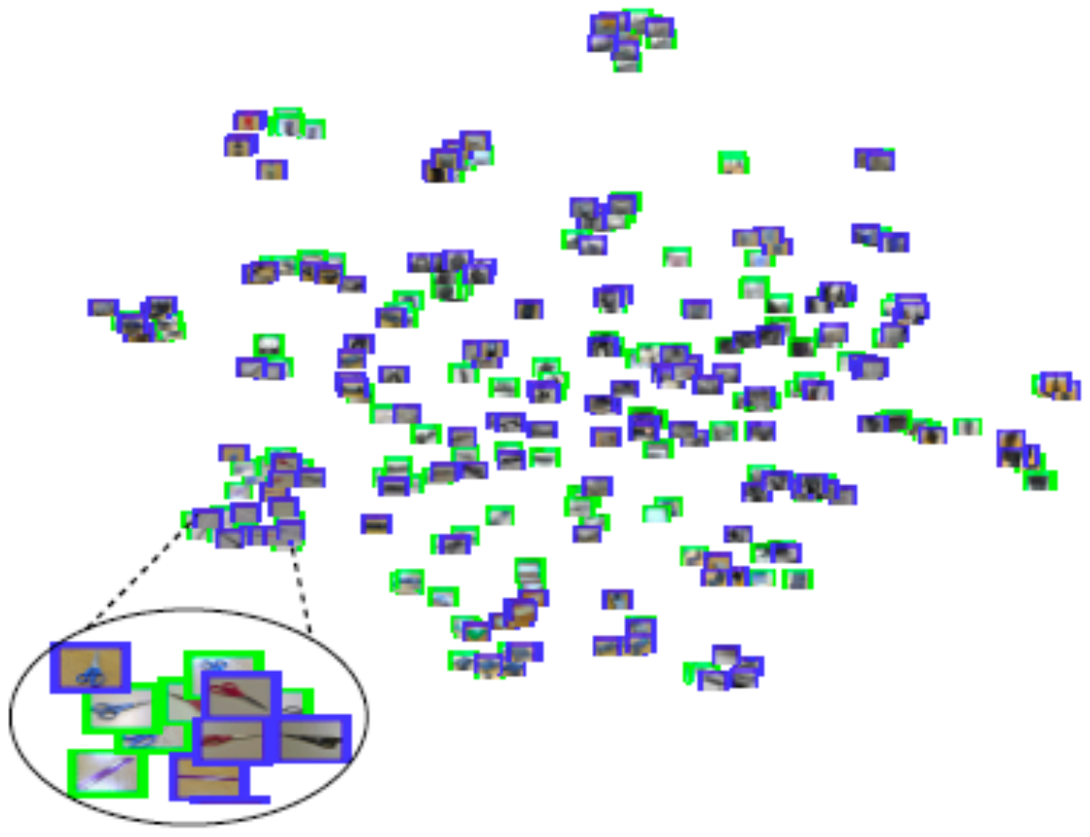
- Multi-class accuracy averaged across 5 train/test splits for domain shift
- Three ways of training
  - with only source data (S)
  - with only target data (T)
  - with source and target data (ST)

Domain Adaptation 2/2

- DeCAF robust to resolution changes (t-SNE algorithm)



(a) SURF features



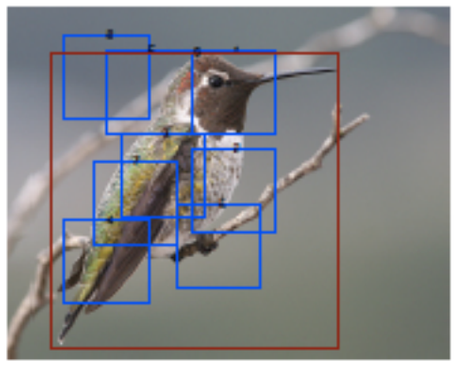
(b) $DeCAF_6$

- DeCAF provides better category clustering than SURF
- DeCAF clusters same category instances across domains

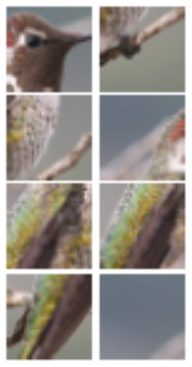
Fine-Grained Recognition (subcategory recognition)

- Caltech-UCSD birds dataset *[Welinder et al., 2010]*
- Performance comparison against several state-of-the-art baselines

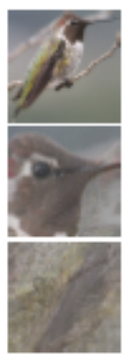
- Two approaches
  - First adopt ImageNet-like pipeline, DeCAF6 and a multi-class logistic regression
  - Second adopt deformable part descriptors (DPD) method *[Zhang et al., 2013]*



(a) DPM detections



(b) Parts



(c) DPD

Method	Accuracy
$DeCAF_6$	58.75
DPD + $DeCAF_6$	**64.96**
DPD (Zhang et al., 2013)	50.98
POOF (Berg & Belhumeur, 2013)	56.78

- Outperforms also POOF with the best accuracy performed in the literature

Scene Recognition

- SUN-397 large-scale scene recognition database *[Xiao et al., 2010]*
- Goal: classify the scene of the entire image

- Used 50 training samples and 50 test samples per class
  - Results averaged across 5 splits of 50 training images and 50 test images
  - Top-performing method selected by cross-validation

	DeCAF_6	DeCAF_7
LogReg	$\mathbf{40.94 \pm 0.3}$	40.84 ± 0.3
SVM	39.36 ± 0.3	40.66 ± 0.3
Xiao et al. (2010)	38.0	

- Outperforms Xiao ed al. (2010), the current state-of-the-art method

- DeCAF demonstrate
  - the ability to generalize to other tasks
  - representational power as compared to traditional hand-engineered features

Discussion

DONE

- Analysis of the use of deep features applied in semi-supervised multi-task framework

DEMONSTRATIONS

- Using a large labeled object database to train a deep convolutional architecture
  - is possible to learn features with representational power and generalization ability
  - is possible to perform good semantic visual discrimination tasks with linear classifiers
  - outperform current state-of-the-art approaches

VISUAL RESULTS

- Demonstrate the generality and semantic knowledge implicit in DeCAF features
- Showing that features tend to cluster images into interesting semantic categories

NUMERICAL RESULTS

- DeCAF frameworks can improve the performance of a wide variety of existing method
- Improving across a spectrum of visual recognition tasks

References

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Multi-task feature learning. In NIPS, 2006.

Berg, A., Deng, J., and Fei-Fei, L. ImageNet large scale visual recognition challenge 2012. 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.

Caruana, R. Multitask learning. Machine Learning, 28, 1997

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In CVPR, 2004.

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

Le, Q., Zou, W., Yeung, S., and Ng, A. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In CVPR, 2011.

Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., and Ng, A. Building high-level features using large scale unsupervised learning. In ICML, 2012.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989.

Li, L., Su, H., Fei-Fei, L., and Xing, E. Object bank: A highlevel image representation for scene classification & semantic feature sparsification. In NIPS, 2010.

Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV, 2001.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In ECCV, 2010.

Singh, S., Gupta, A., and Efros, A. Unsupervised discovery of mid-level discriminative patches. In ECCV, 2012.

Torresani, L., Szummer, M., and Fitzgibbon, A. Efficient object category recognition using classemes. In ECCV. 2010.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. JMLR, 9, 2008.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. Locality-constrained linear coding for image classification. In CVPR, 2010.

Welinder, P., Branson, S., Mita, T.,Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, 2010.

Yang, J., L., Y., Tian, Y., Duan, L., and Gao, W. Group-sensitive multiple kernel learning for object categorization.ICCV, 2009

Zhang, N., Farrell, R., Iandola, F., and Darrell, T. Deformable part descriptors for fine-grained recognition and attribute prediction. In ICCV, 2013.

Some links

- DeCAF release (GitHub) : https://github.com/UCB-ICSI-Vision-Group/decaf-release/
- Caffe (DeCAF improvement) : http://caffe.berkeleyvision.org/
- Alex Krizhevsky convolutional neural network : https://code.google.com/p/cuda-convnet/
- ILSVRC-2012 : http://www.image-net.org/challenges/LSVRC/2012/
- ImageNet database : http://www.image-net.org/
- t-SNE: http://homepage.tudelft.nl/19j49/t-SNE.html