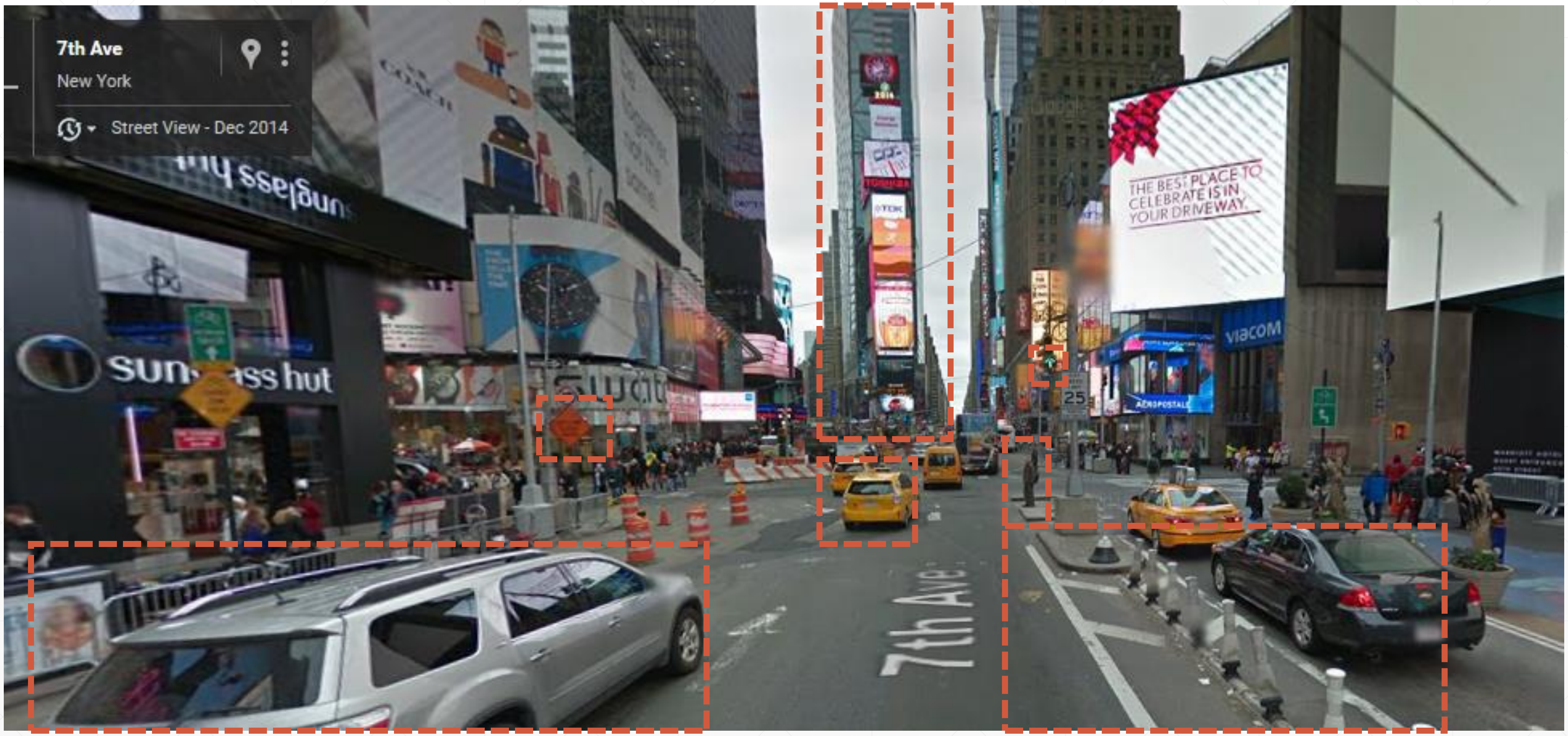
DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition

ECE 289G: Paper Presentation #3

Philipp Gysel

Autonomous Car

Real-time object recognition



Source: maps.google.com

Object classification

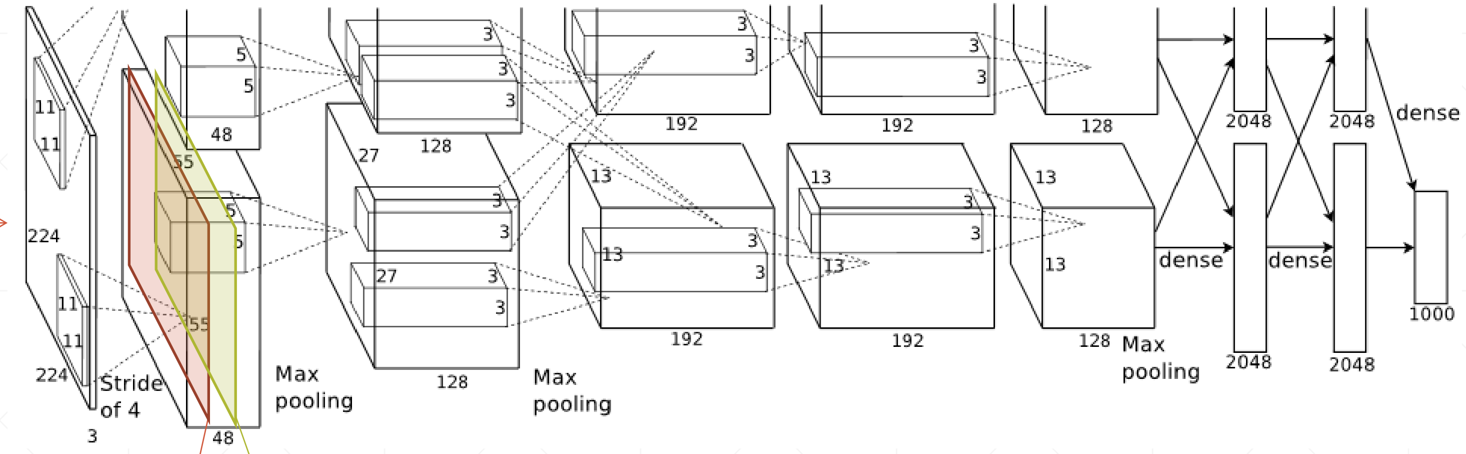
Convolutional Neural Network

Training

- Car
- Traffic Light
- Street Sign
- …

Source: imagenet.stanford.ed

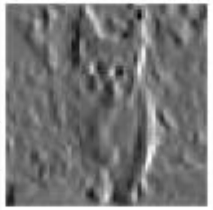
CNN for Object Recognition



Object
category

Feature extraction

Classification

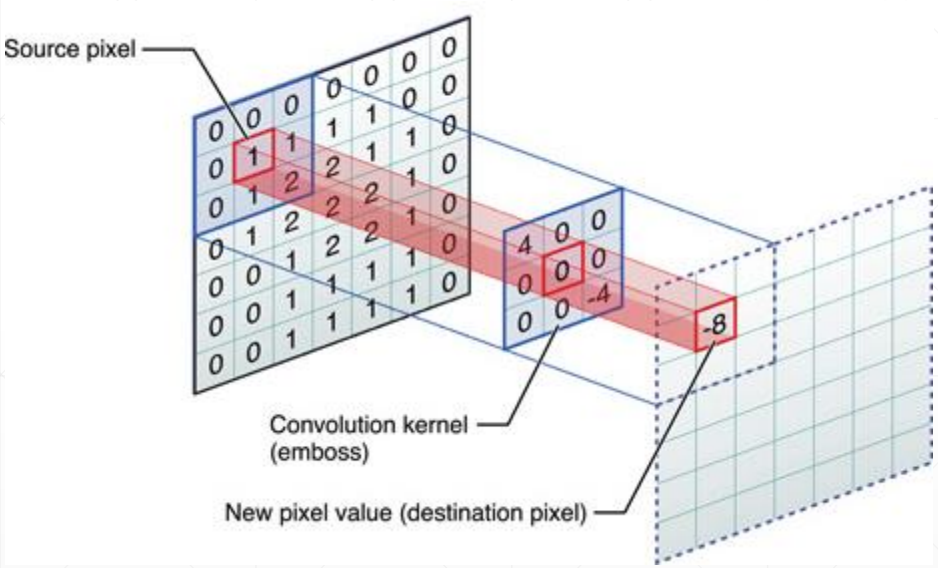


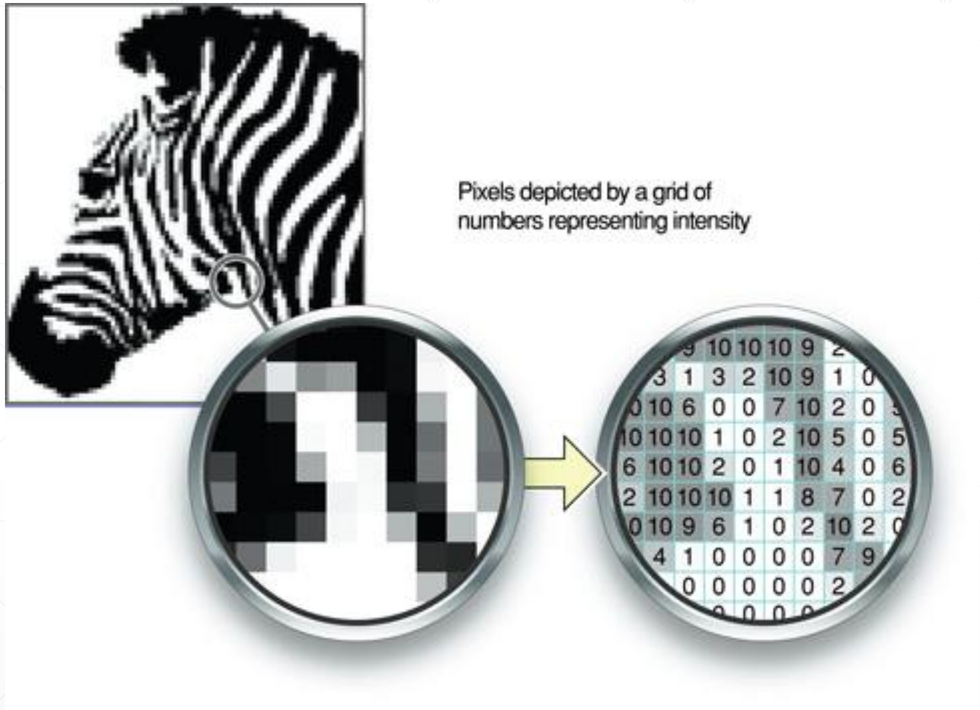
ectangles
Gradients

Leave
Building

Source: [2]

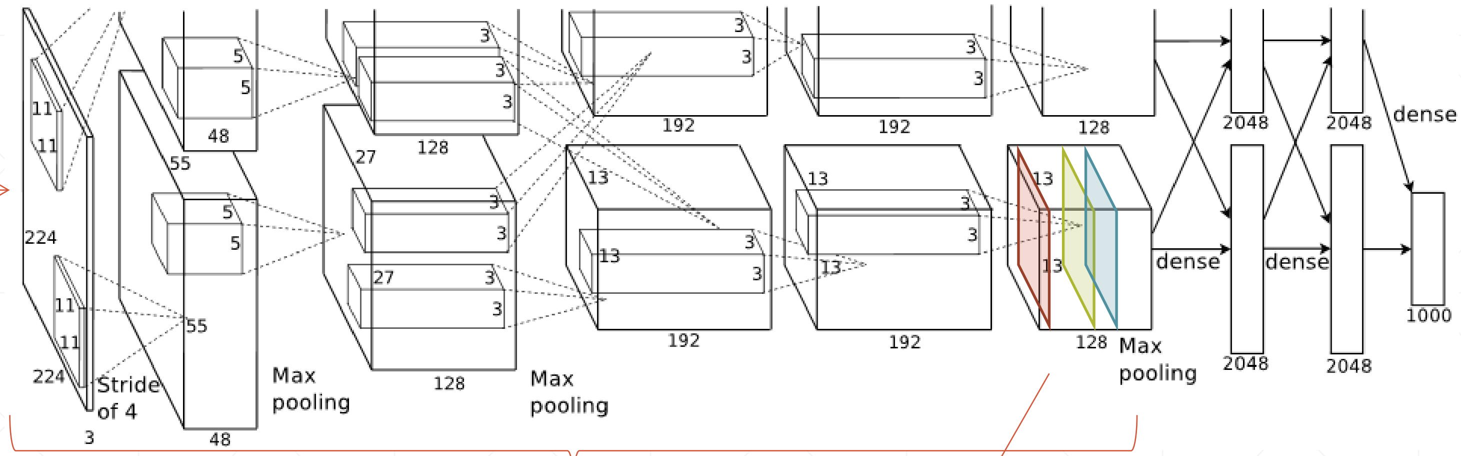
Feature extraction





Source: https://developer.apple.com

From features to object classes



Feature extraction

Object category

High-level features:
- Shape of a car
- Road marking
- Face with eyes and ears
- Cat skin

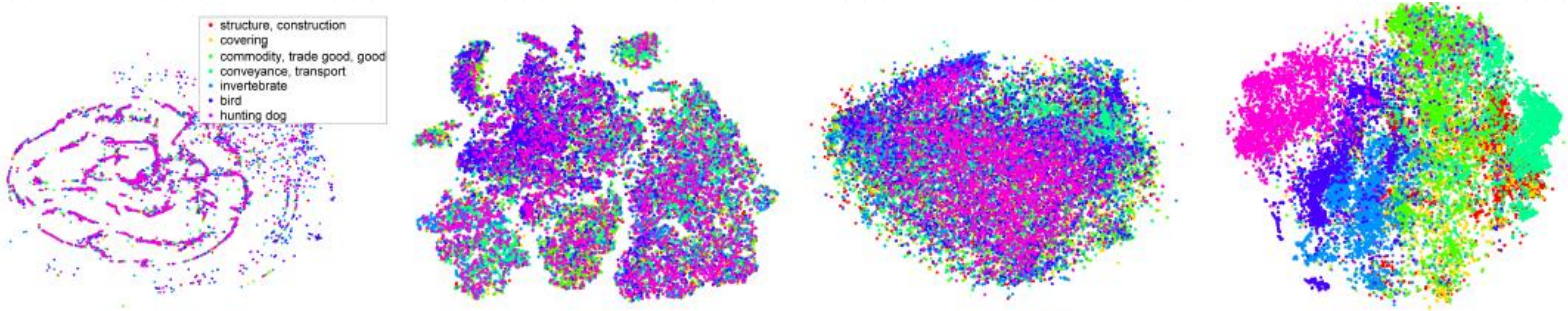
Classification

Classes:
- Cat
- Car
- …

Source: [2]

Visualization of high dimensional feature space

- LLC [5] vs GIST [6] vs DeCAF [1]
- Vizualisation with t-SNE algorithm [4]



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

Source: [1]

Repurpose Features from CNN

→ Convolutional Neural Network → Object class

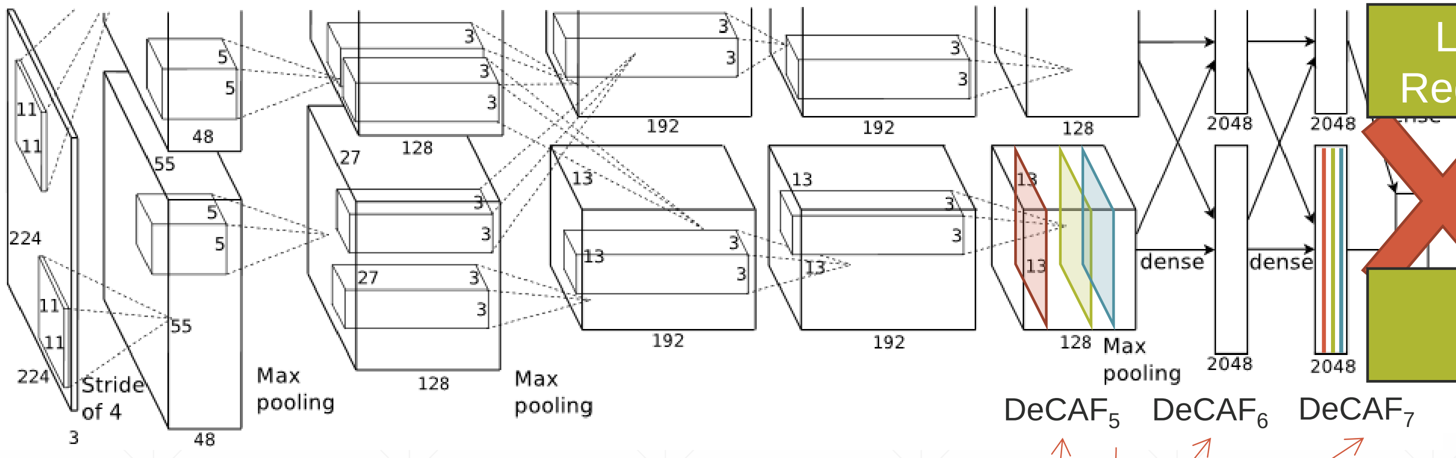
↑ Learned Features

→ Convolutional Neural Network

Source: imagenet.stanford.ed

Classification with small training dataset

Target database



Logistic Regression

SVM

Classify new database

Freeze trained convolution kernels

High-level features

Source: [2]

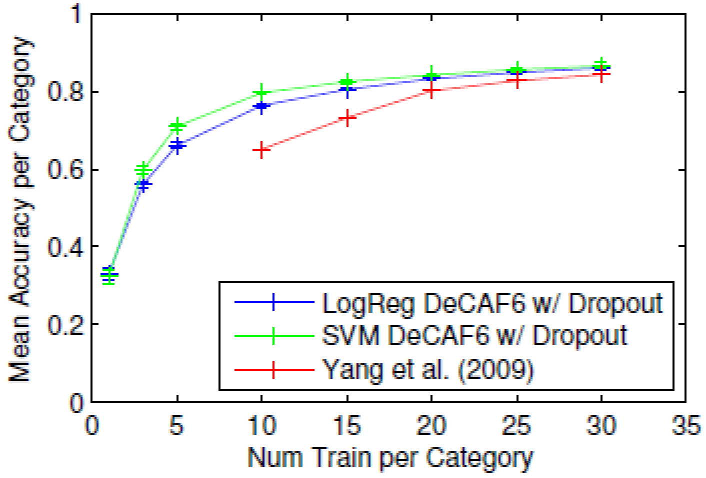
Experiments: Are features transferrable to solve new tasks?

- Train AlexNet [2] on ILSVRC 2012 object recognition dataset
- Reuse extracted features for new tasks:
- Experiment #1: Basic Object Recognition
- Experiment #2: Domain Adaption
- Experiment #3: Fine-grained recognition
- Experiment #4: Scene recognition

Experiment #1: Basic object recognition

- Classify new objects on new dataset (Caltech-101 dataset)
- 2.6% better than state-of-art

	$DeCAF_5$	$DeCAF_6$	$DeCAF_7$
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	



Source: [1]

Experiment #2: Domain adaption

- Train object recognition in different surrounding, only few labeled data in target domain available
- Office dataset

	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	94.79 ± 1.2	92.96 ± 2.0
Daume III (2007)	40.26 ± 1.1	82.14 ± 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	

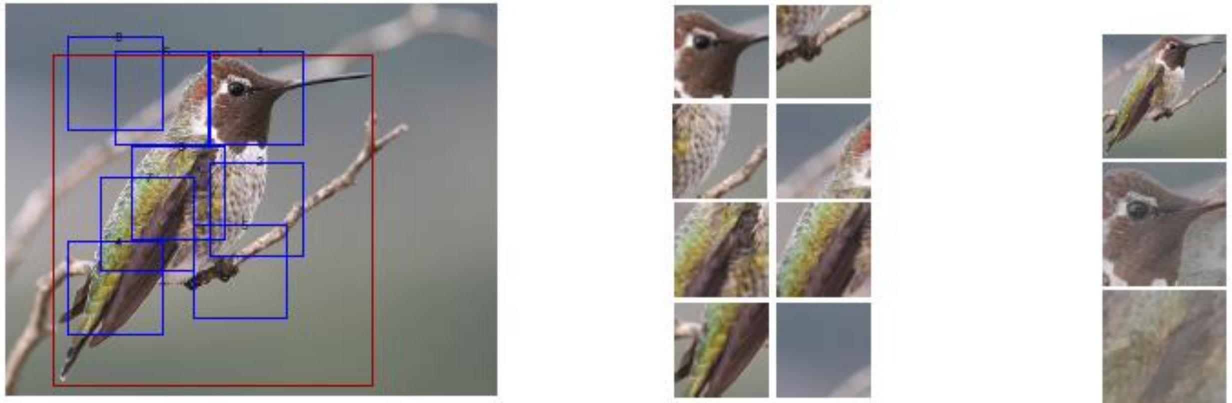
Source: [1]

Experiment #3: Subcategory recognition

- Caltech-UCSD birds dataset
- 8% better than state-of-art

Method	Accuracy
DeCAF$_6$	58.75
DPD + DeCAF$_6$	**64.96**
DPD (Zhang et al., 2013)	50.98
POOF (Berg & Belhumeur, 2013)	56.78

Table 2. Accuracy on the Caltech-UCSD bird dataset.



(a) DPM detections (b) Parts (c) DPD

Source: [1]

Experiment #4: Scene recognition

- Classes like abbey, diner, mosque, stadium
- SUN-397 dataset
- >2% better than state-of-art

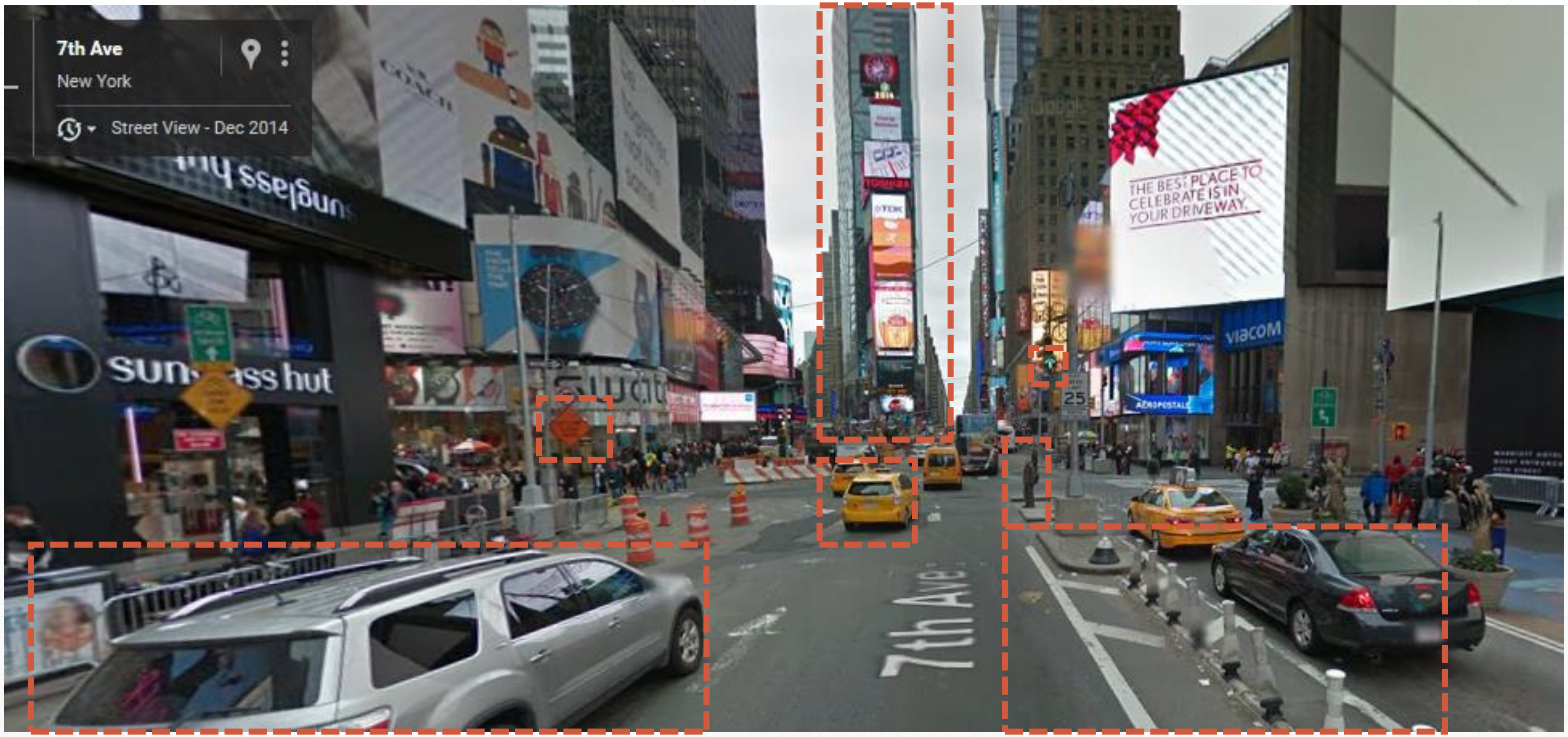
	DeCAF_6	DeCAF_7
LogReg	$\mathbf{40.94 \pm 0.3}$	40.84 ± 0.3
SVM	39.36 ± 0.3	40.66 ± 0.3
Xiao et al. (2010)	38.0	

Source: [1]

Conclusions

- Extract features from ILSVRC dataset to solve new classification tasks
- State-of-the-art performance in 4 different tasks
- CNN features are generic enough to solve completely new problems
- Bigger datasets yield better accuracy
- Release of DeCAF (predecessor of Caffe)

Conclusions cont.



Source: maps.google.com

Conclusions cont.

Convolutional Neural Network

- Car
- Traffic Light
- Street Sign
- …

Training

- Challenges:
  - Find labeled data
  - Training time of CNN

Source: imagenet.stanford.ed

Q&A

References

[1] Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition." *arXiv preprint arXiv:1310.1531* (2013).

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

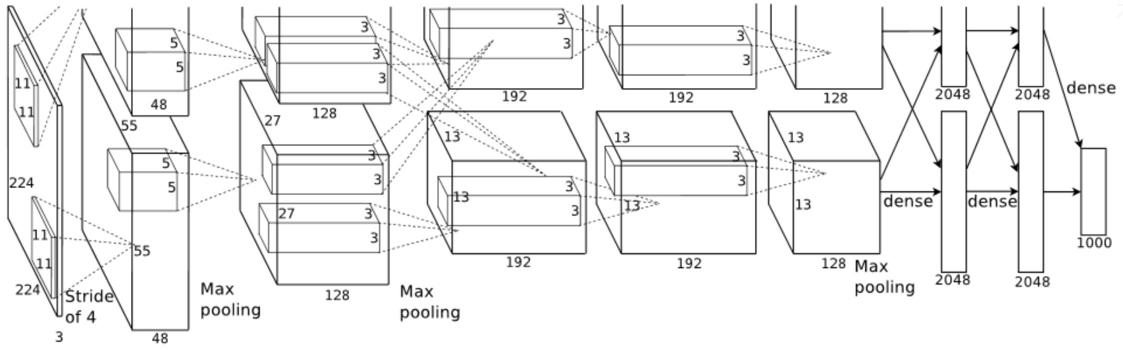
[3] Chopra, S., Balakrishnan, S., and Gopalan, R. Dlid: Deep learning for domain adaptation by interpolating between domains. In ICML Workshop on Challenges in Representation Learning, 2013.

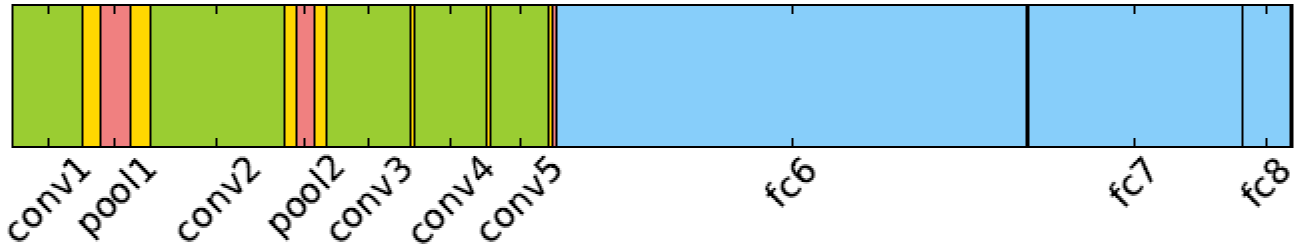
[4] van der Maaten, L. and Hinton, G. Visualizing data using t-sne. JMLR, 9, 2008.

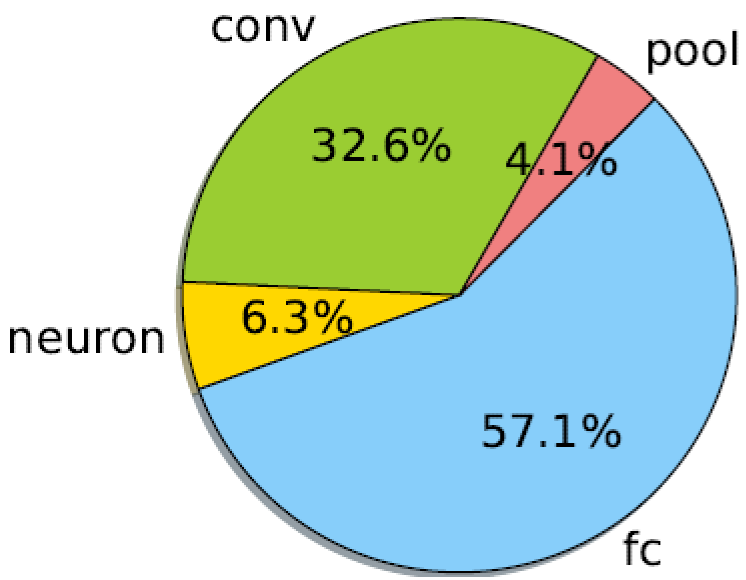
[5] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. Locality-constrained linear coding for image classification. In CVPR, 2010.

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

Computing time of forward propagation







Source: [1] and [2]