Directions in Convolutional Neural Networks at Google

Jon Shlens
Google Research
2 March 2015

Goals

- Provide a broad (and *incomplete*) survey of vision research applying deep networks at Google
- Avoid details but describing overview of problem.
- Almost all of the work I did not do. My amazing colleagues did it.

The computer vision competition: IMAGENET

Large scale academic competition focused on predicting 1000 object classes (~1.2M images).

...
electric ray, crampfish, numbfish, torpedo
sawfish
smalltooth sawfish, Pristis pectinatus
guitarfish
stingray
roughtail stingray, Dasyatis centroura
...

. . .

Imagenet: A large-scale hierarchical image database
J Deng et al (2009)

History of techniques in ImageNet Challenge

ImageNet 2010

Locality constrained linear coding + SVM	NEC & UIUC
Fisher kernel + SVM	Xerox Research Center Europe
SIFT features + LI2C	Nanyang Technological Institute
SIFT features + k-Nearest Neighbors	Laboratoire d'Informatique de Grenoble
Color features + canonical correlation analysis	National Institute of Informatics, Tokyo

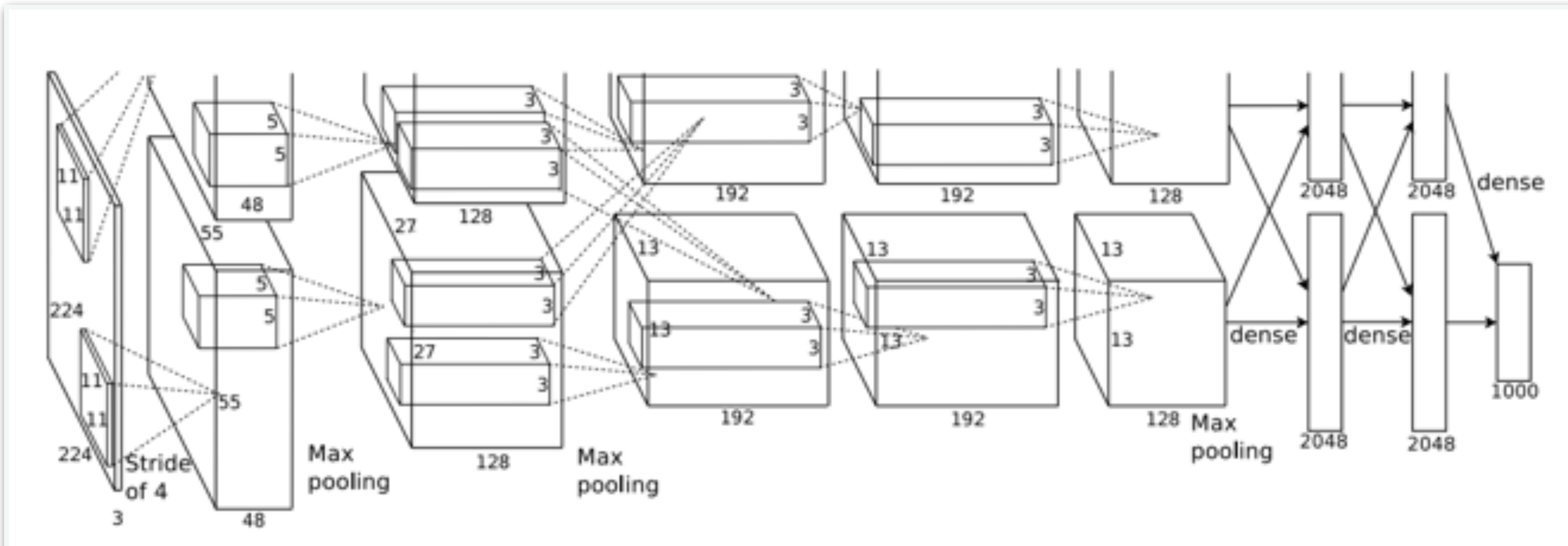
ImageNet 2011

Compressed Fisher kernel + SVM	Xerox Research Center Europe
SIFT bag-of-words + VQ + SVM	University of Amsterdam & University of
SIFT + ?	ISI Lab, Tokyo University

ImageNet 2012

Deep convolutional neural network	University of Toronto
Discriminatively trained DPMs	University of Oxford
Fisher-based SIFT features + SVM	ISI Lab, Tokyo University

Convolutional neural networks, revisited



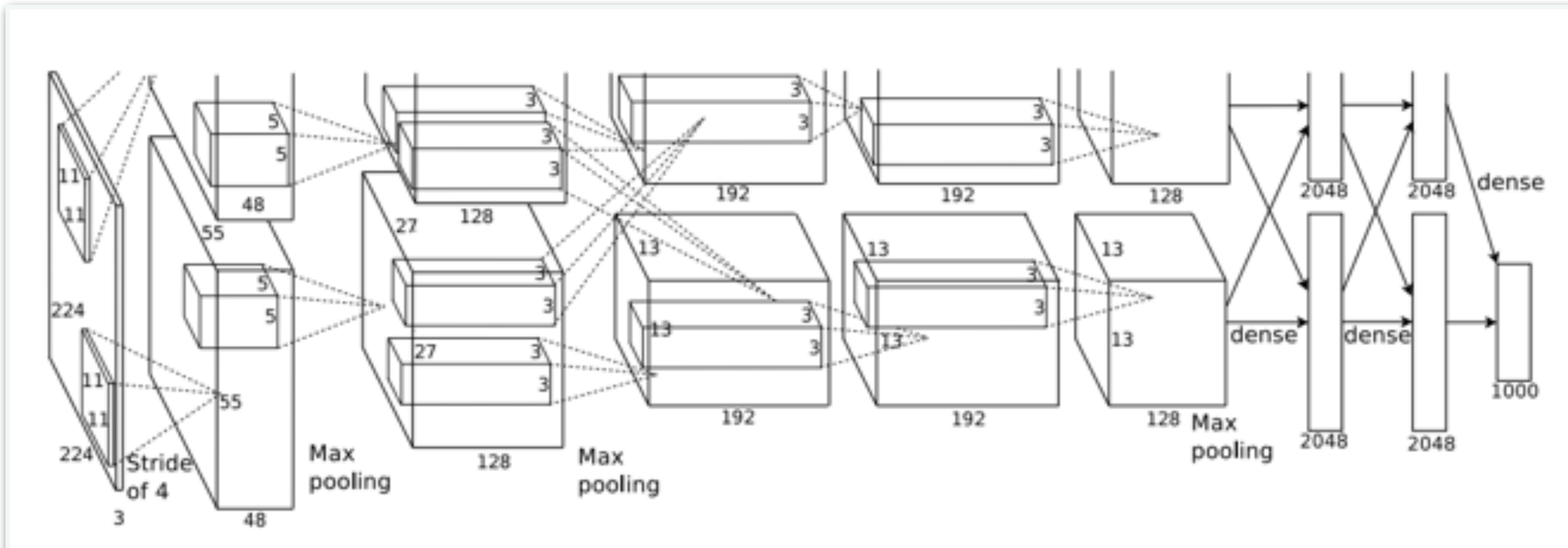
ImageNet Classification with Deep Convolutional Neural Networks
A Krizhevsky I Sutskever, G Hinton (2012)

- Repeated motifs of convolution, local response normalization and max pooling across ~13 layers.

- Most elements of network architecture employed as early as the late 1980’s.

Backpropagation applied to handwritten zip code recognition
Y LeCun et al (1990)

What happened?



ImageNet Classification with Deep Convolutional Neural Networks
A Krizhevsky I Sutskever, G Hinton (2012)

- Winning network contained 60M parameters.
- Achieving <u>scale</u> in compute and data is critical.
    - large academic data sets
    - SIMD hardware (e.g. GPU's, SSE instruction sets)

Applications at Google (and beyond)

- Image Search
- Image Labeling
- Image Segmentation
- Object Detection
- Object Tracking
- Photo OCR
- Video Annotation
- Video Recommendation
- Fine-grained Classification
- Robot Perception
- Microscopy Analysis

Outline

- Architectures for building vision models — Dist-Belief, Inception
- New methods for optimization — batch normalization, adversarial training
- Combining vision with language — DeViSE, Show-And-Tell
- Beyond image recognition — DRAW, video

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- Beyond image recognition

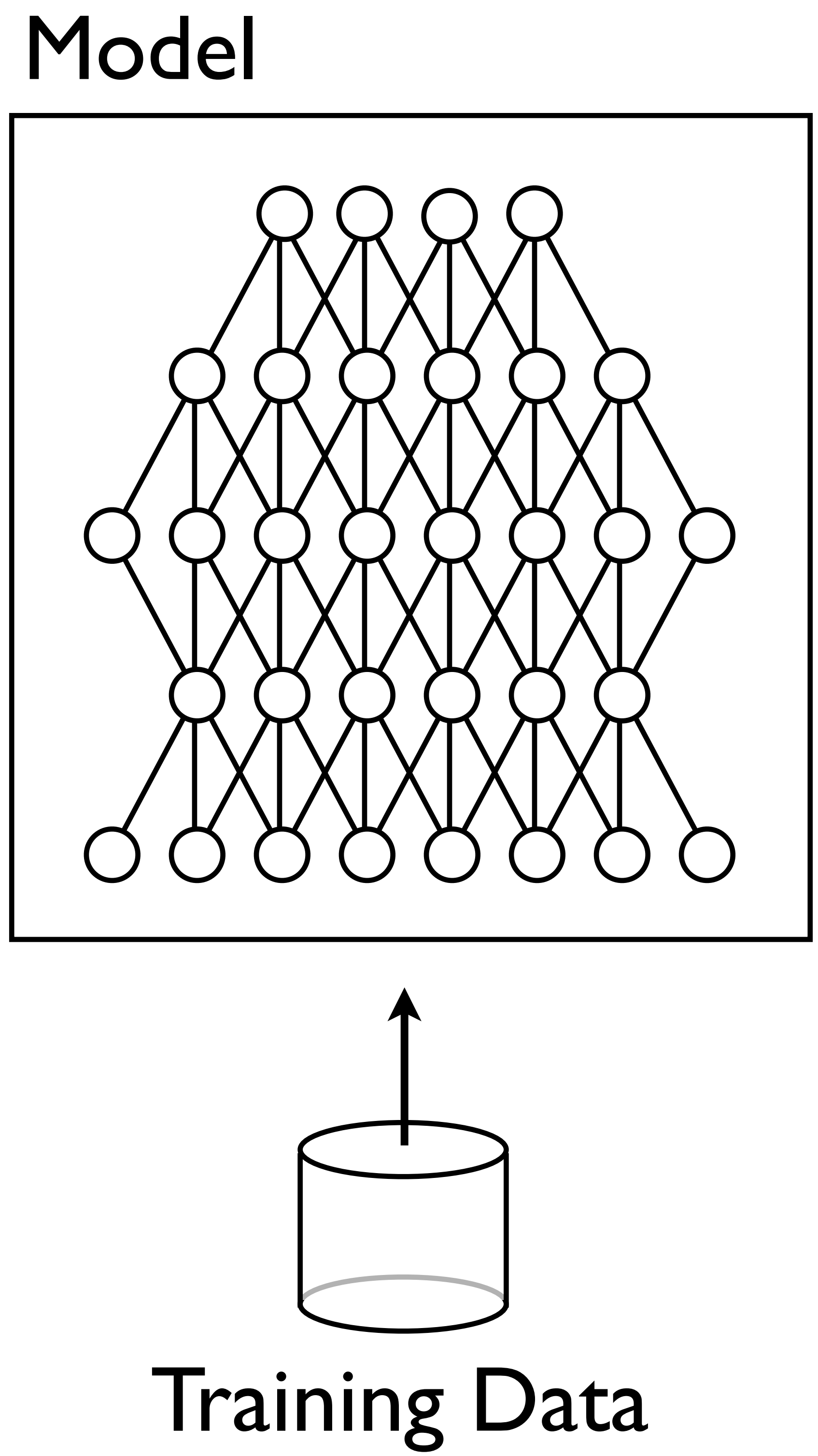
Dist-Belief
Inception

batch normalization
adversarial training

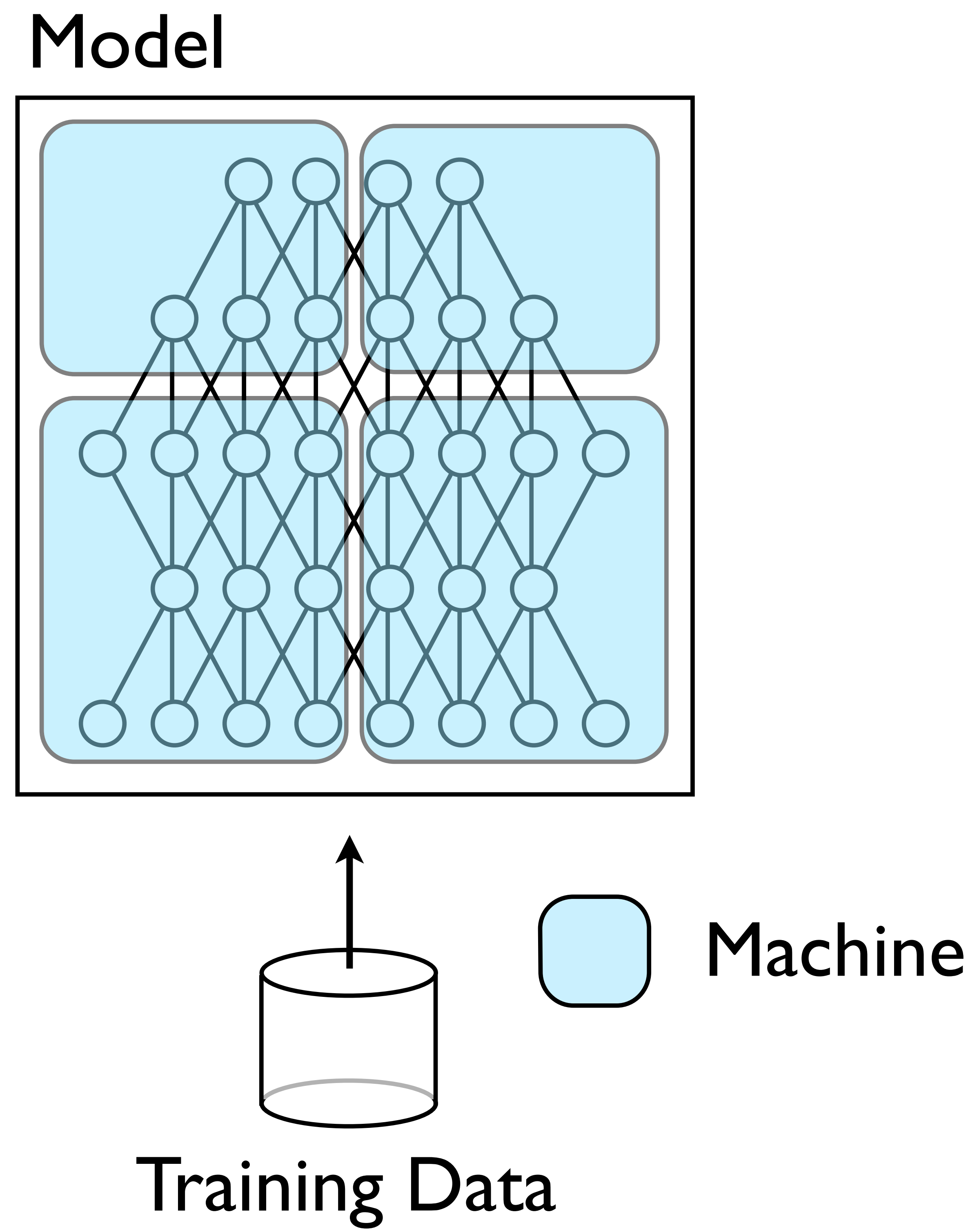
DeViSE
Show-And-Tell

DRAW
video

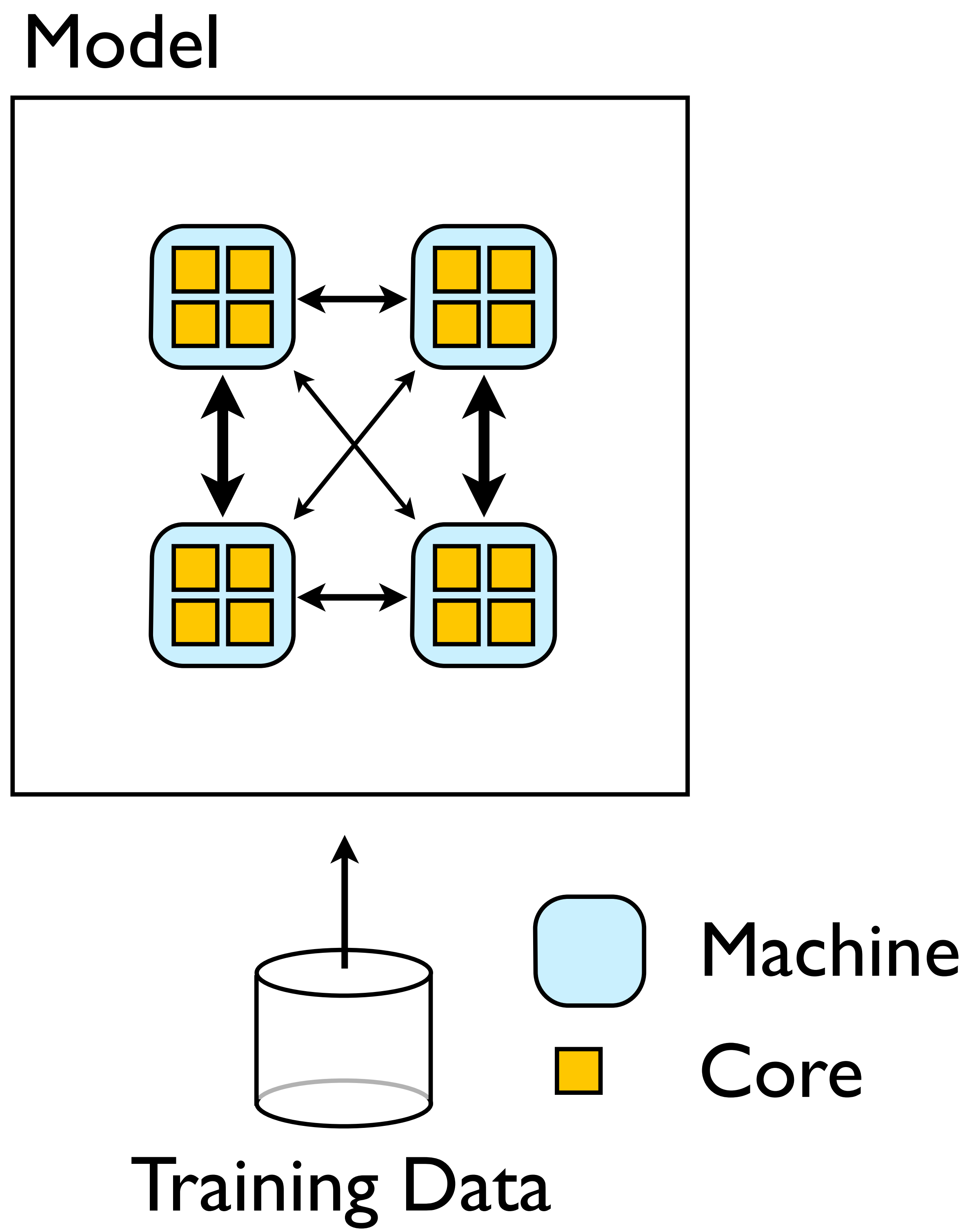
One method to achieve scale is parallelization



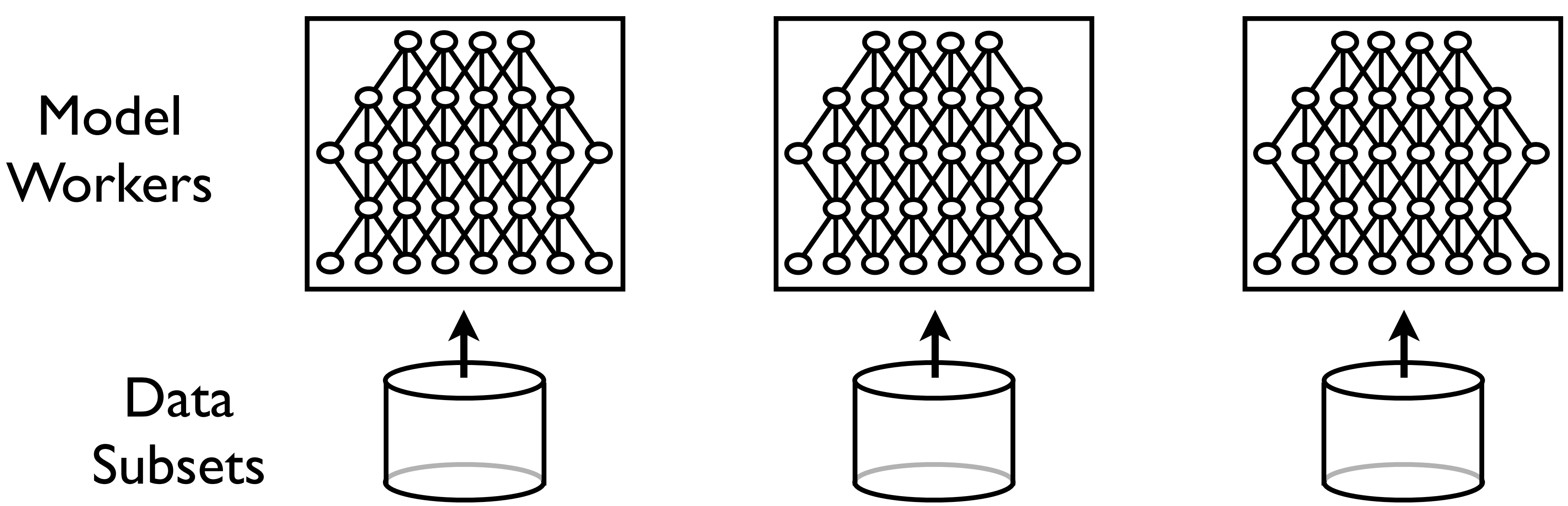
One method to achieve scale is parallelization



One method to achieve scale is parallelization

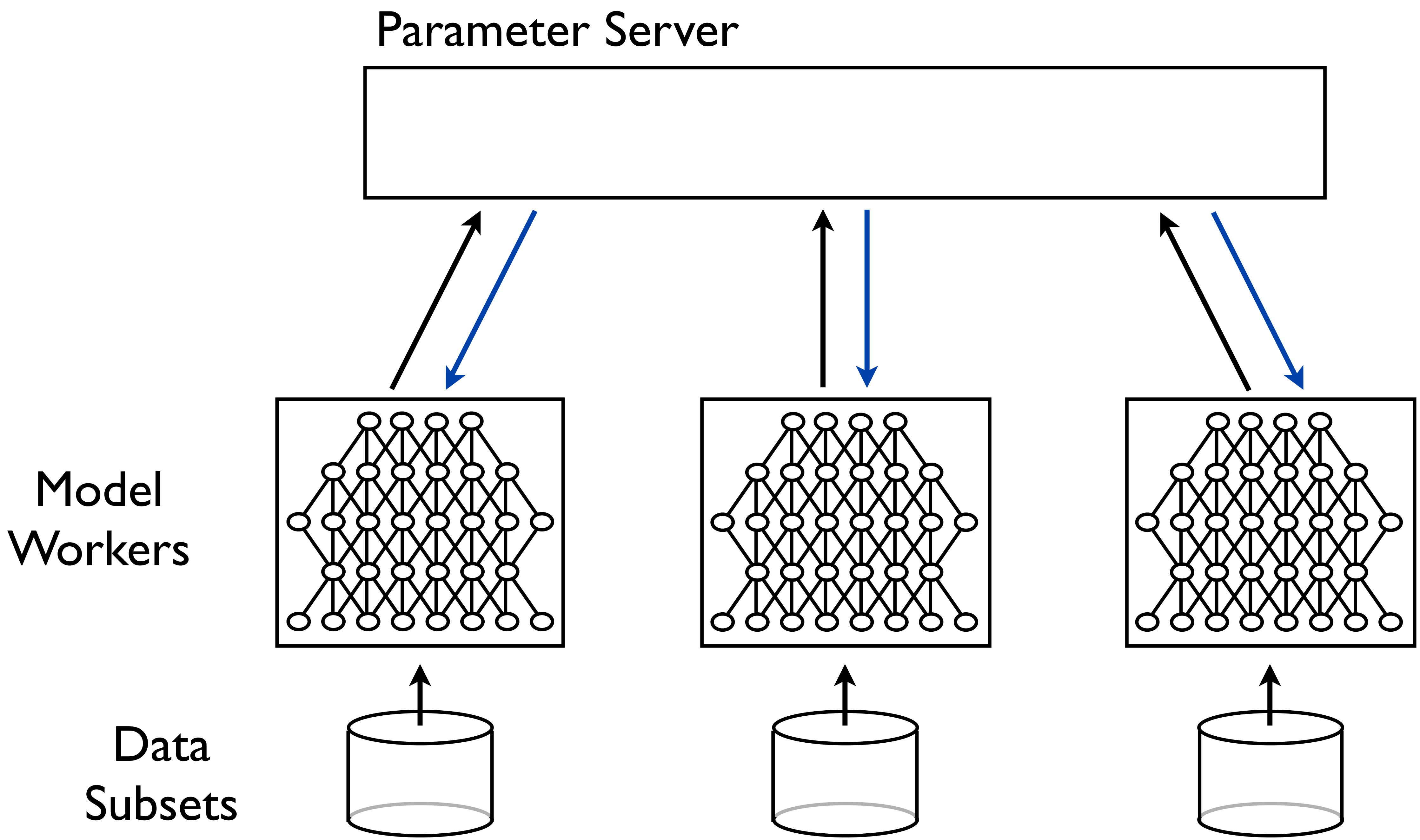


One method to achieve scale is parallelization



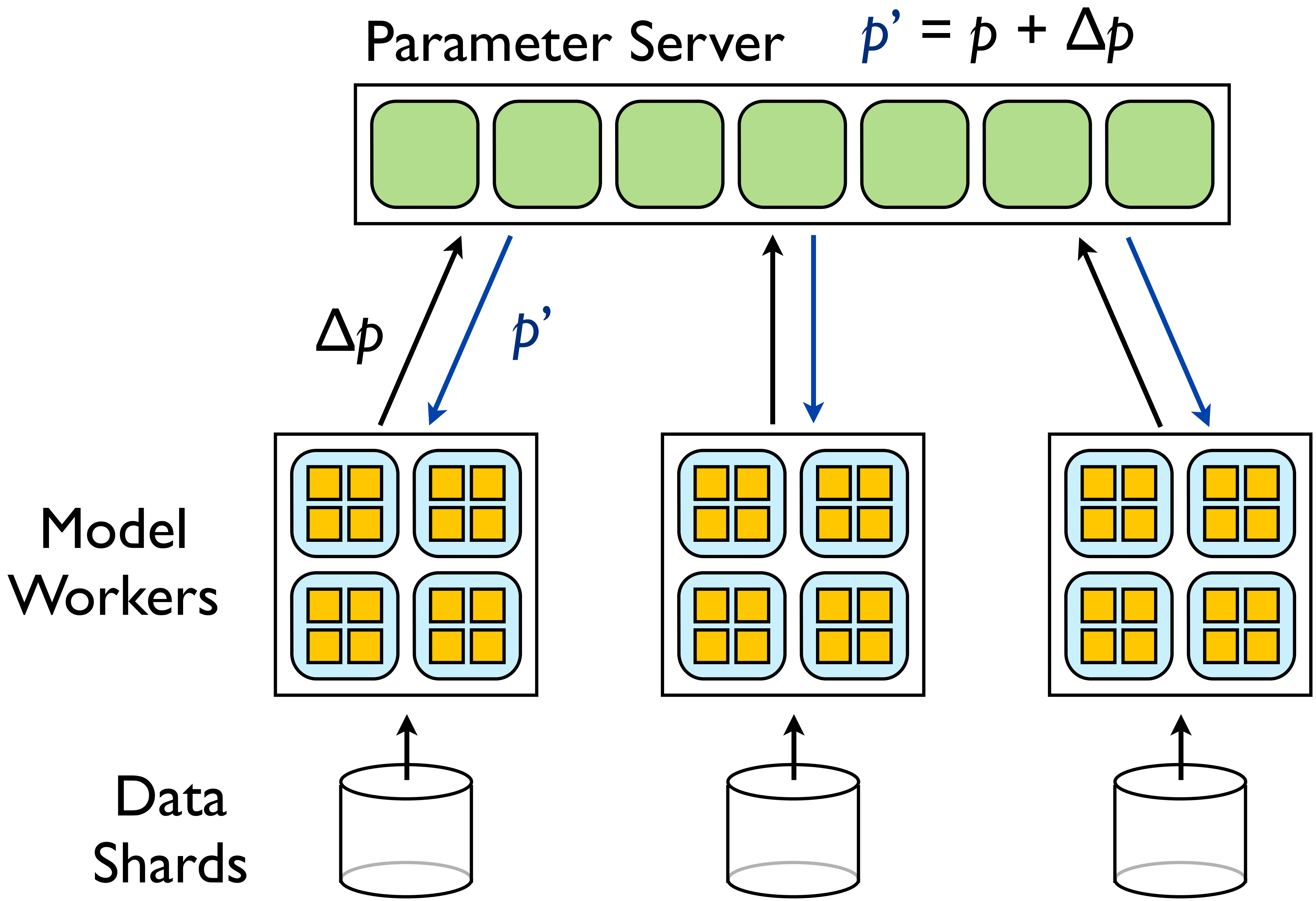
Large scale distributed deep networks
J Dean et al (2012)

One method to achieve scale is parallelization



Large scale distributed deep networks
J Dean et al (2012)

One method to achieve scale is parallelization



Large scale distributed deep networks
J Dean et al (2012)

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- New directions.

Dist-Belief
Inception

batch normalization
adversarial training

DeViSE
Show-And-Tell

DRAW
video

Steady advances in vision architectures.

- Successive improvements to CNN architectures provide steady improvement in image recognition.

		top 5 error
2012	Krizhevsky, Suskever and Hinton *	16.4%
2013	Zeiler and Fergus *	11.5%
2014	Szegedy et al *	6.6%
2015	He et al	4.9%
2015	Ioffe and Szegedy	4.8%

* winner of ImageNet Challenge

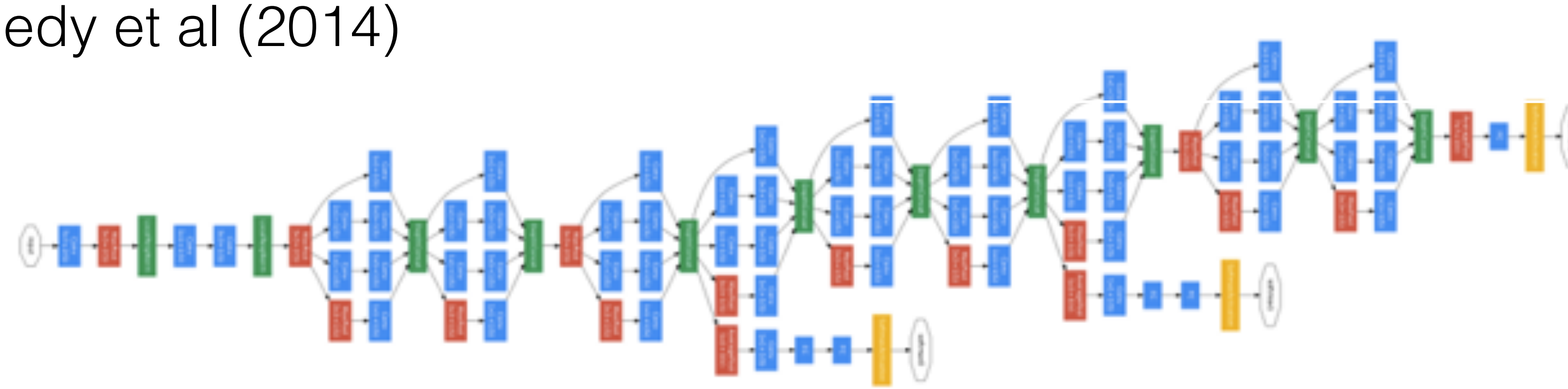
Inception is both better and more efficient.

	params	*FLOPs*
Krizhevsky, Suskever and Hinton (2012)	60M	2B
Zeiler and Fergus (2013)	75M	2B+
Szegedy et al (2014)	5M	1.5B

Krizhevsky, Suskever and Hinton (2012)

Zeiler and Fergus (2013)

Szegedy et al (2014)



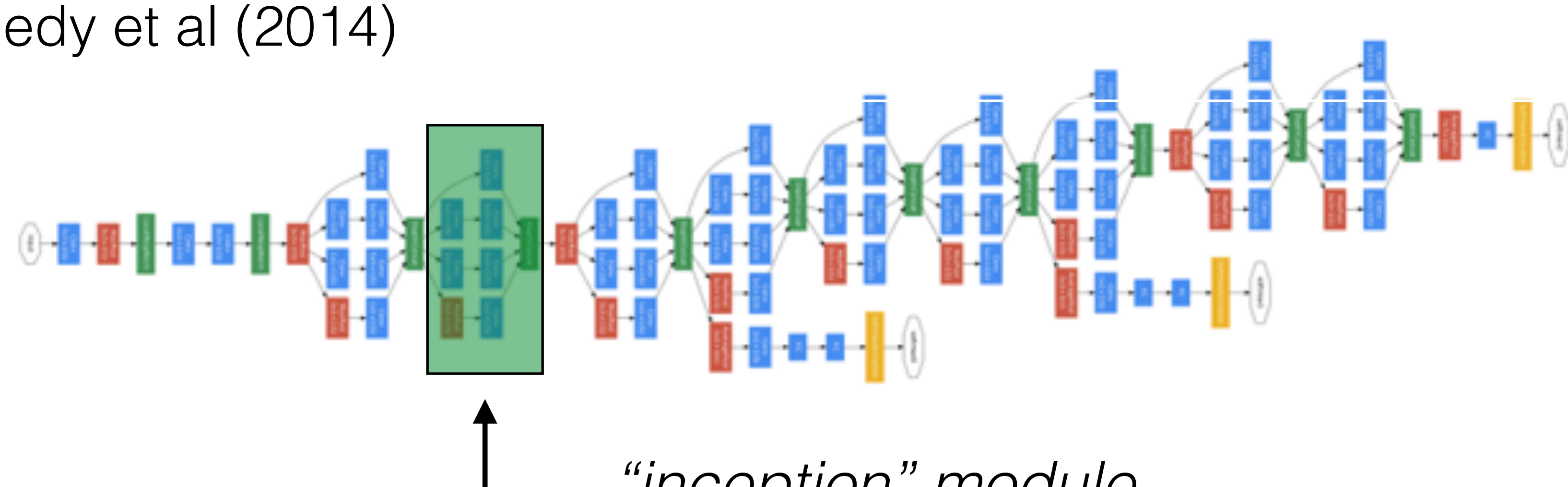
Inception is both better and more efficient.

	params	*FLOPs*
Krizhevsky, Suskever and Hinton (2012)	60M	2B
Zeiler and Fergus (2013)	75M	2B+
Szegedy et al (2014)	5M	1.5B

Krizhevsky, Suskever and Hinton (2012)

Zeiler and Fergus (2013)

Szegedy et al (2014)



"inception" module

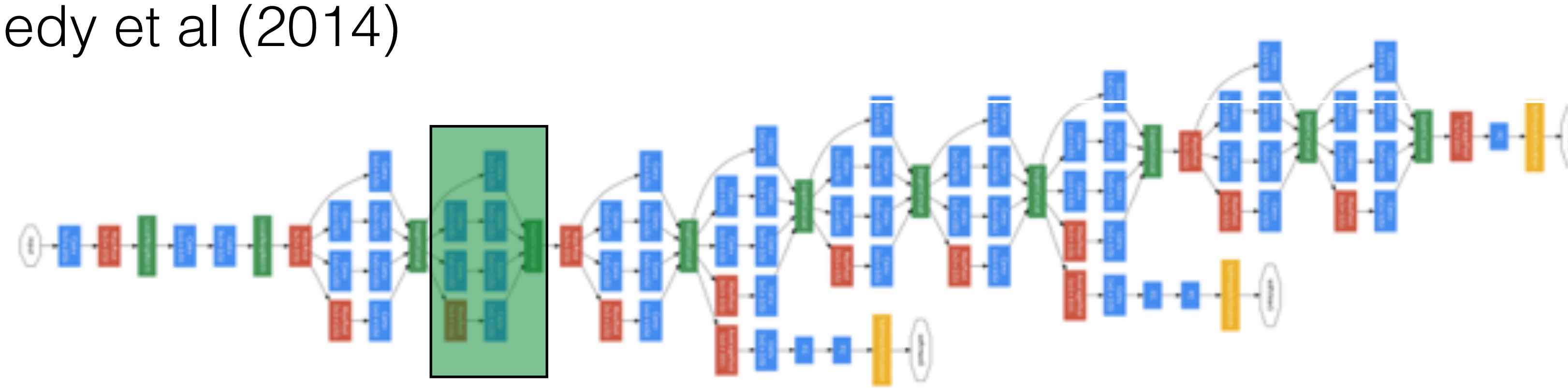
Inception is both better and more efficient.

	params	*FLOPs*
Krizhevsky, Suskever and Hinton (2012)	60M	2B
Zeiler and Fergus (2013)	75M	2B+
Szegedy et al (2014)	5M	1.5B

Krizhevsky, Suskever and Hinton (2012)

Zeiler and Fergus (2013)

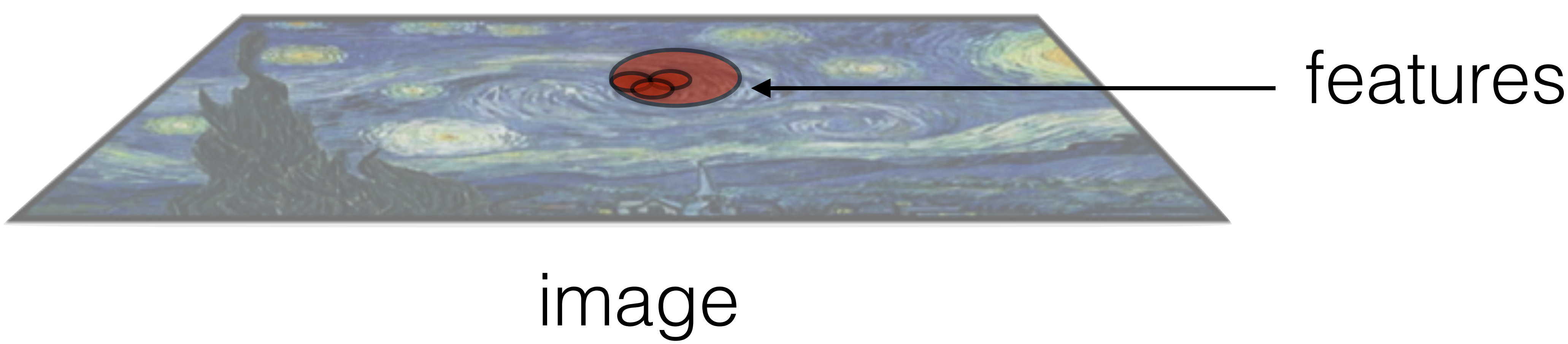
Szegedy et al (2014)



"inception" module

image

Natural images are locally heavily correlated.



Natural Image Statistics and Neural Representation
E Simoncelli and B Olshausen (2001)

Filter activations reflect image correlations

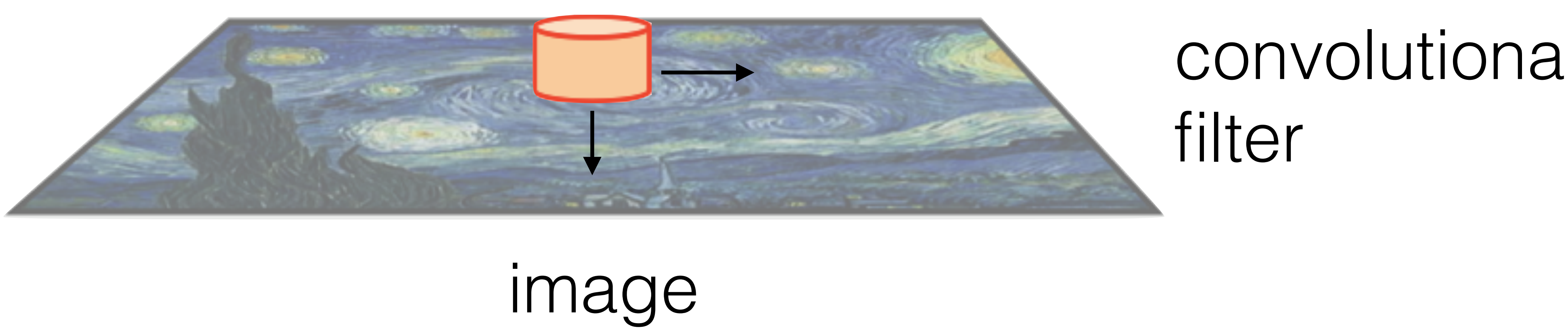
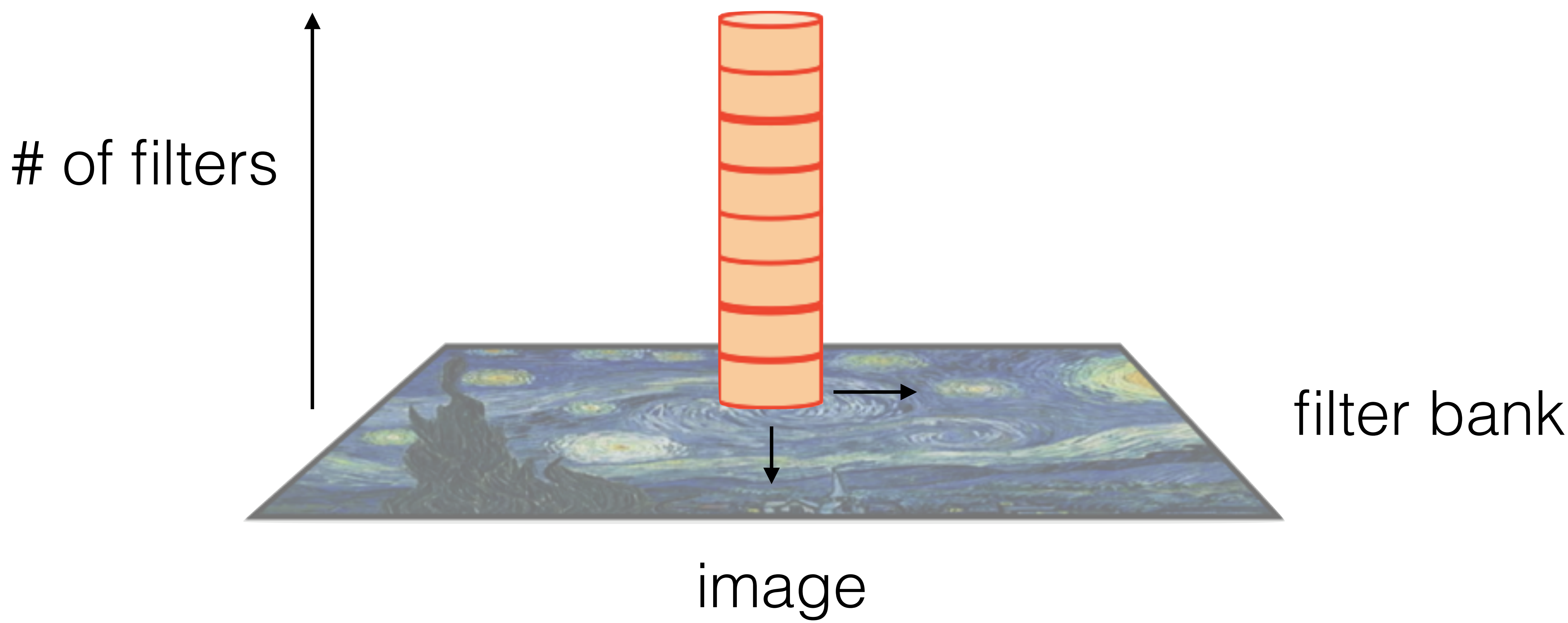
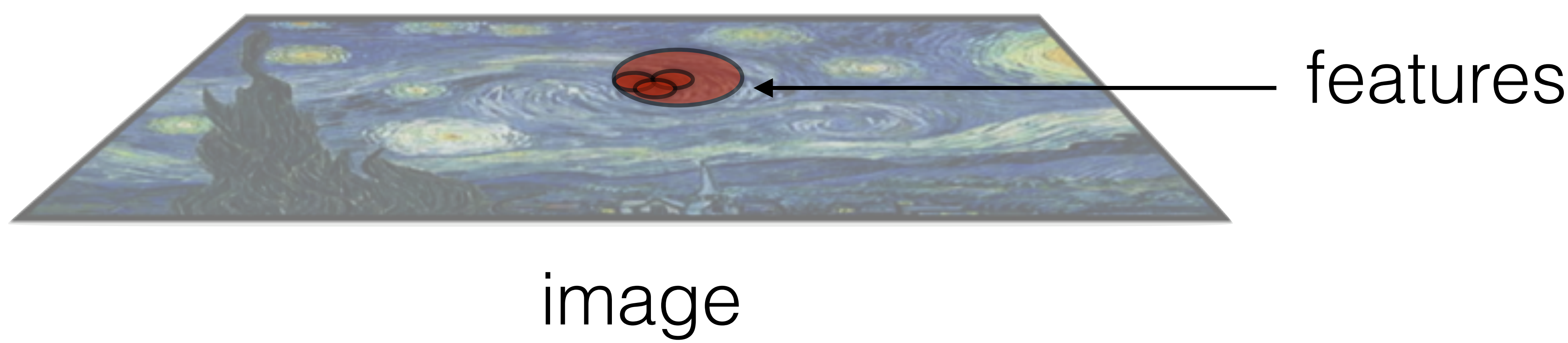


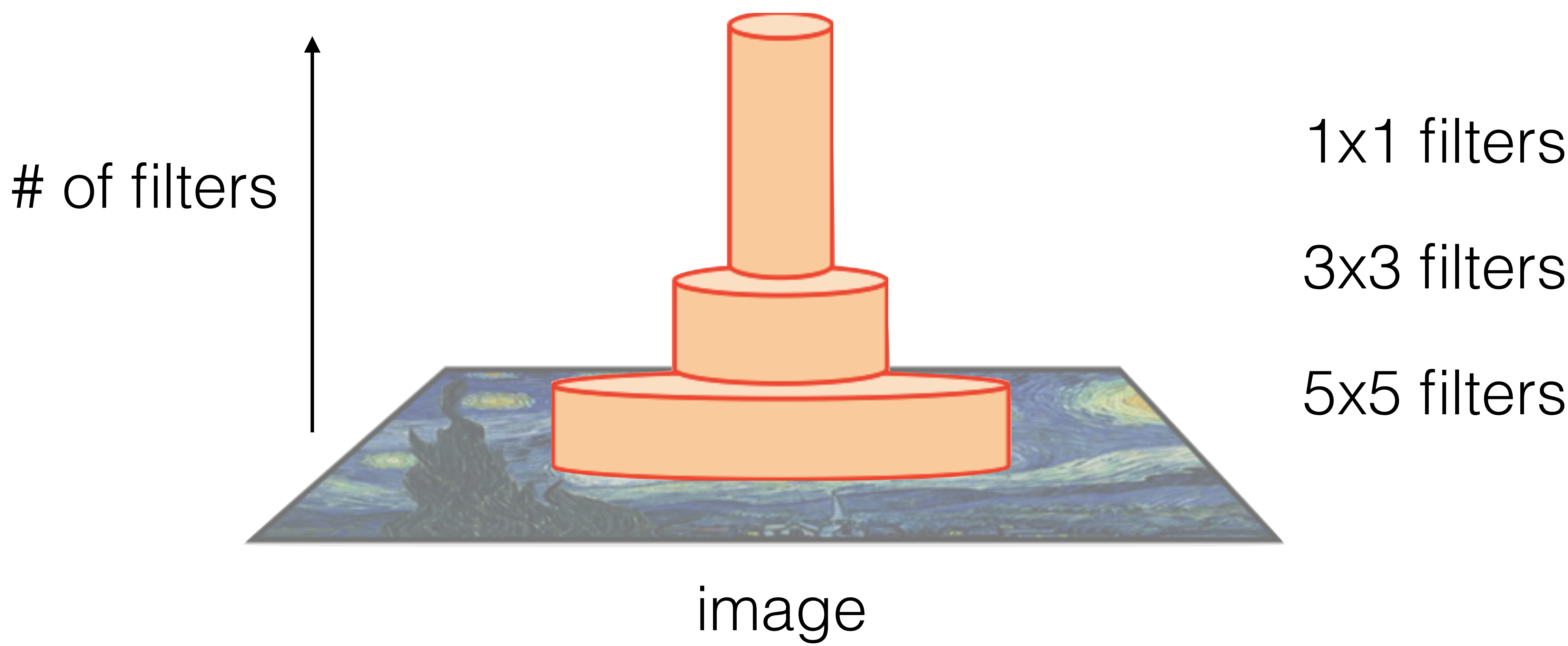
Image correlations reflected in filter bank correlations



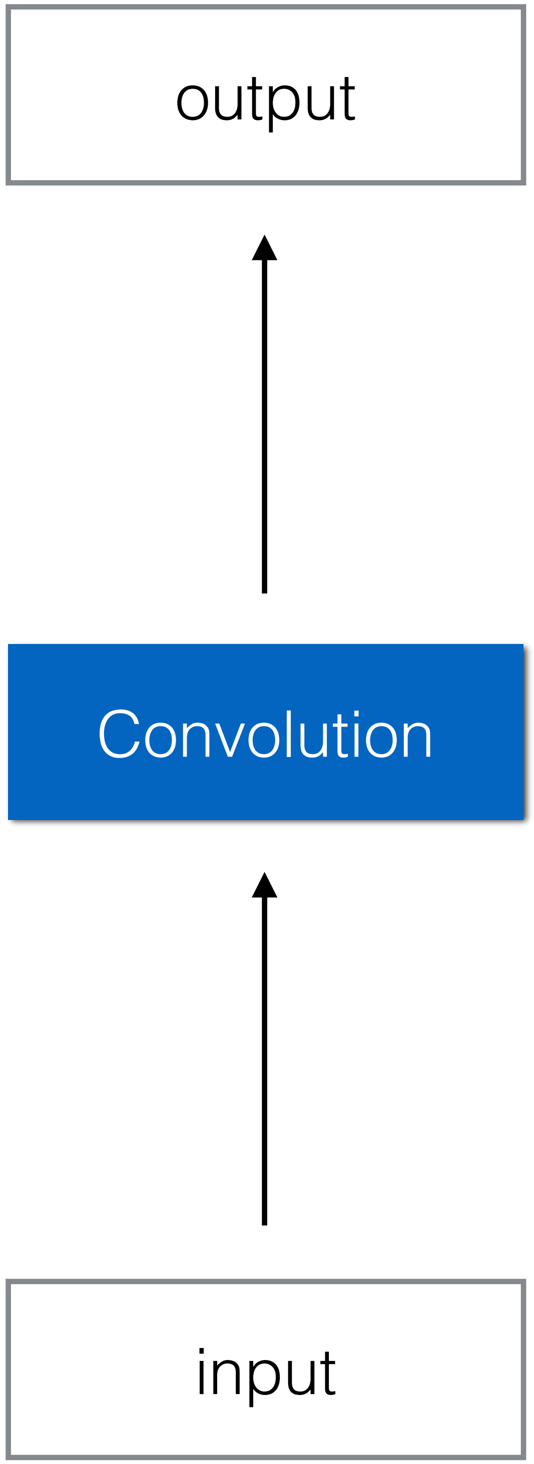
Correlations in natural images are multi-scale



Correlations in natural images are multi-scale

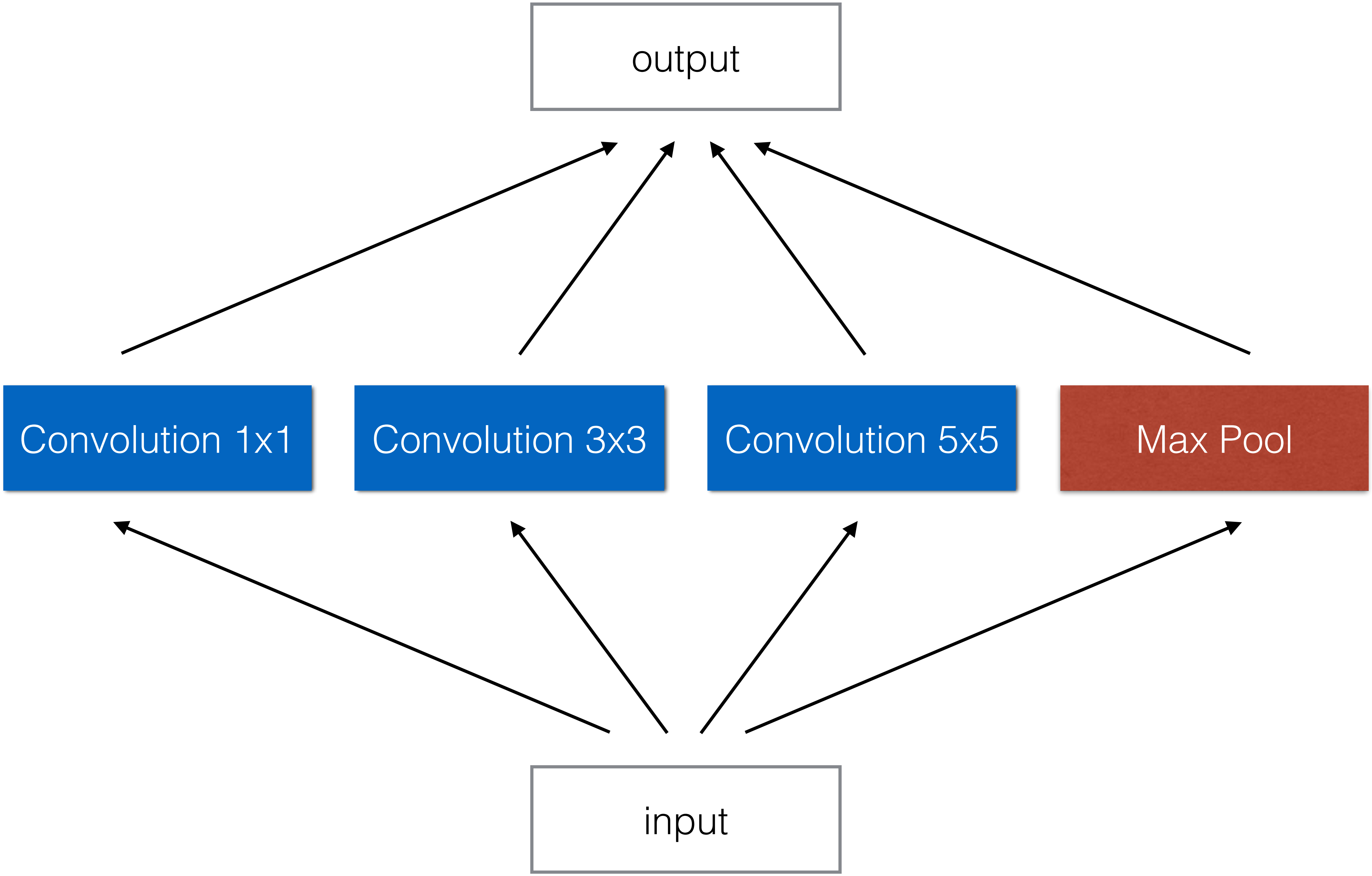


Going Deeper with Convolutions
C Szegedy et al (2014)



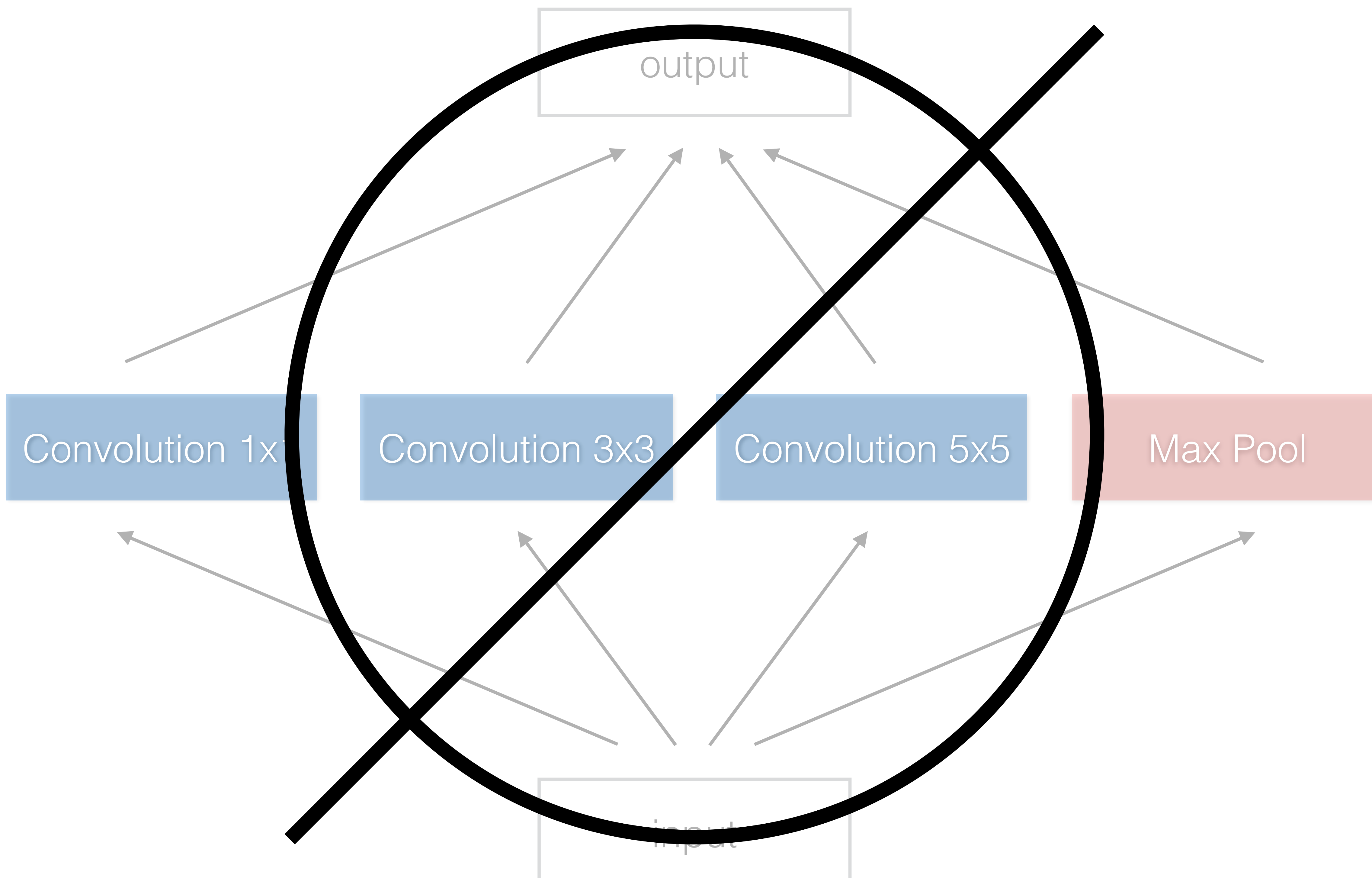
Going Deeper with Convolutions
C Szegedy et al (2014)

Replace convolution with multi-scale convolution



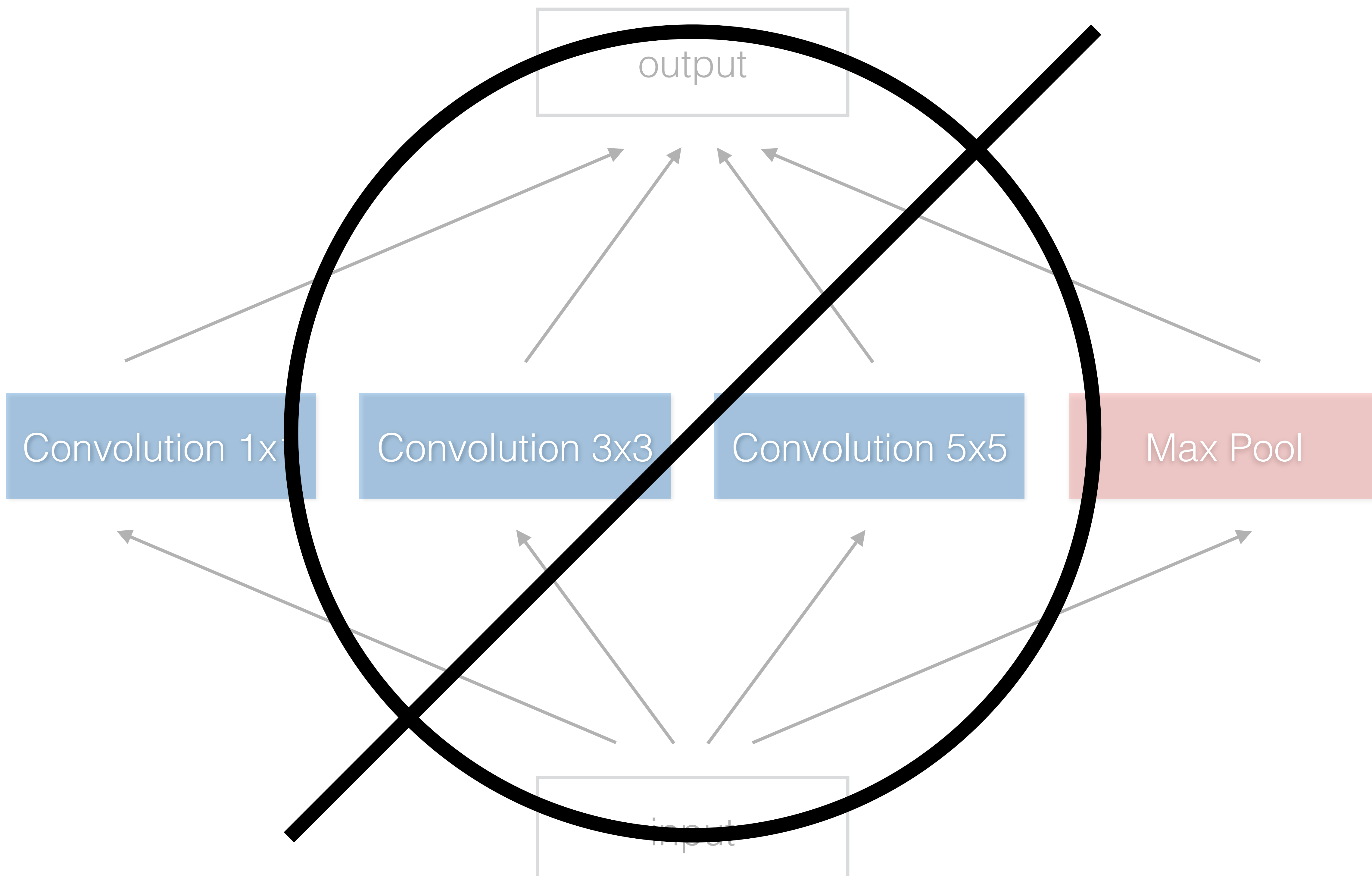
Going Deeper with Convolutions
C Szegedy et al (2014)

Multi-scale representation is *not* sufficient.



Going Deeper with Convolutions
C Szegedy et al (2014)

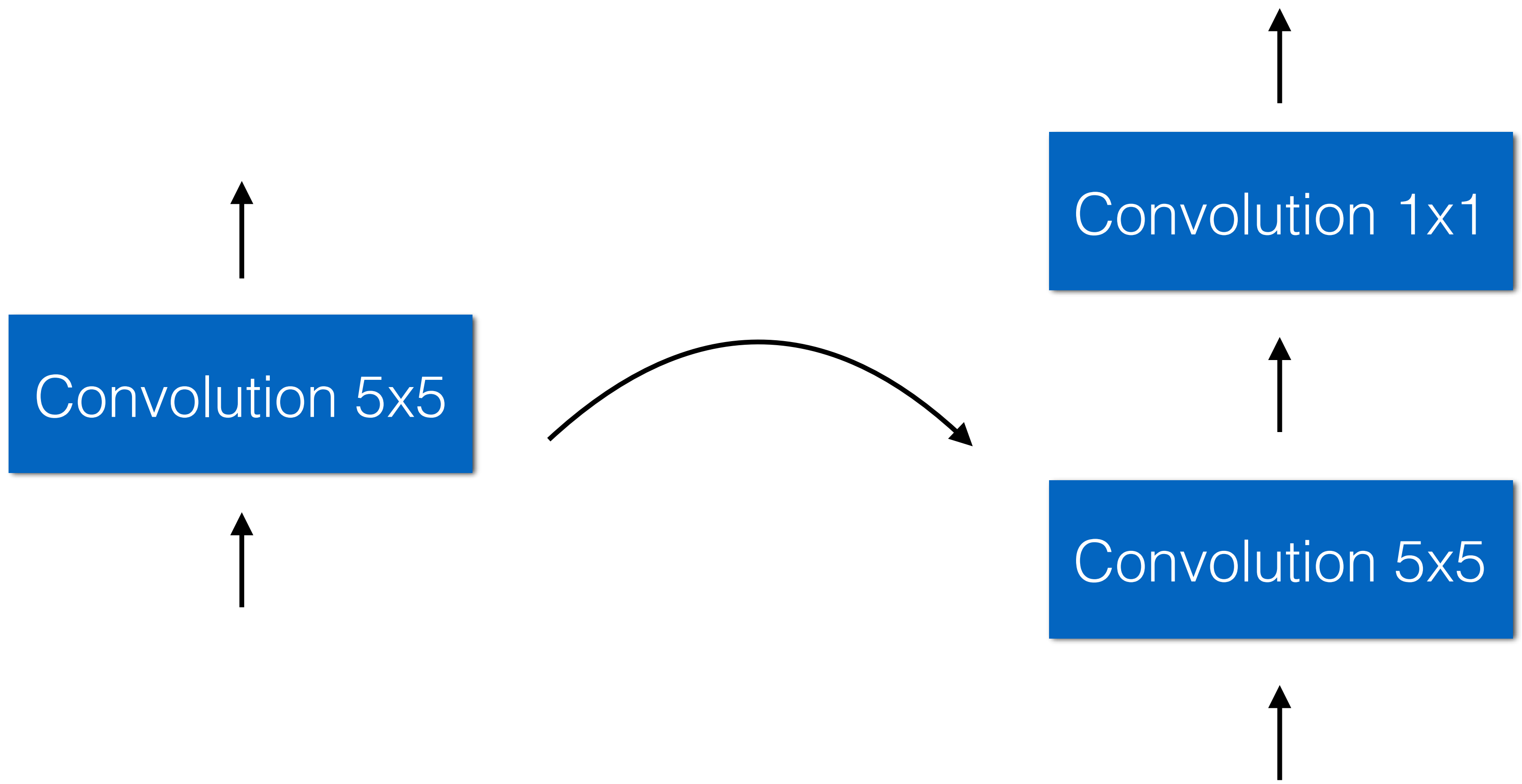
Multi-scale representation is *not* sufficient.



Going Deeper with Convolutions
C Szegedy et al (2014)

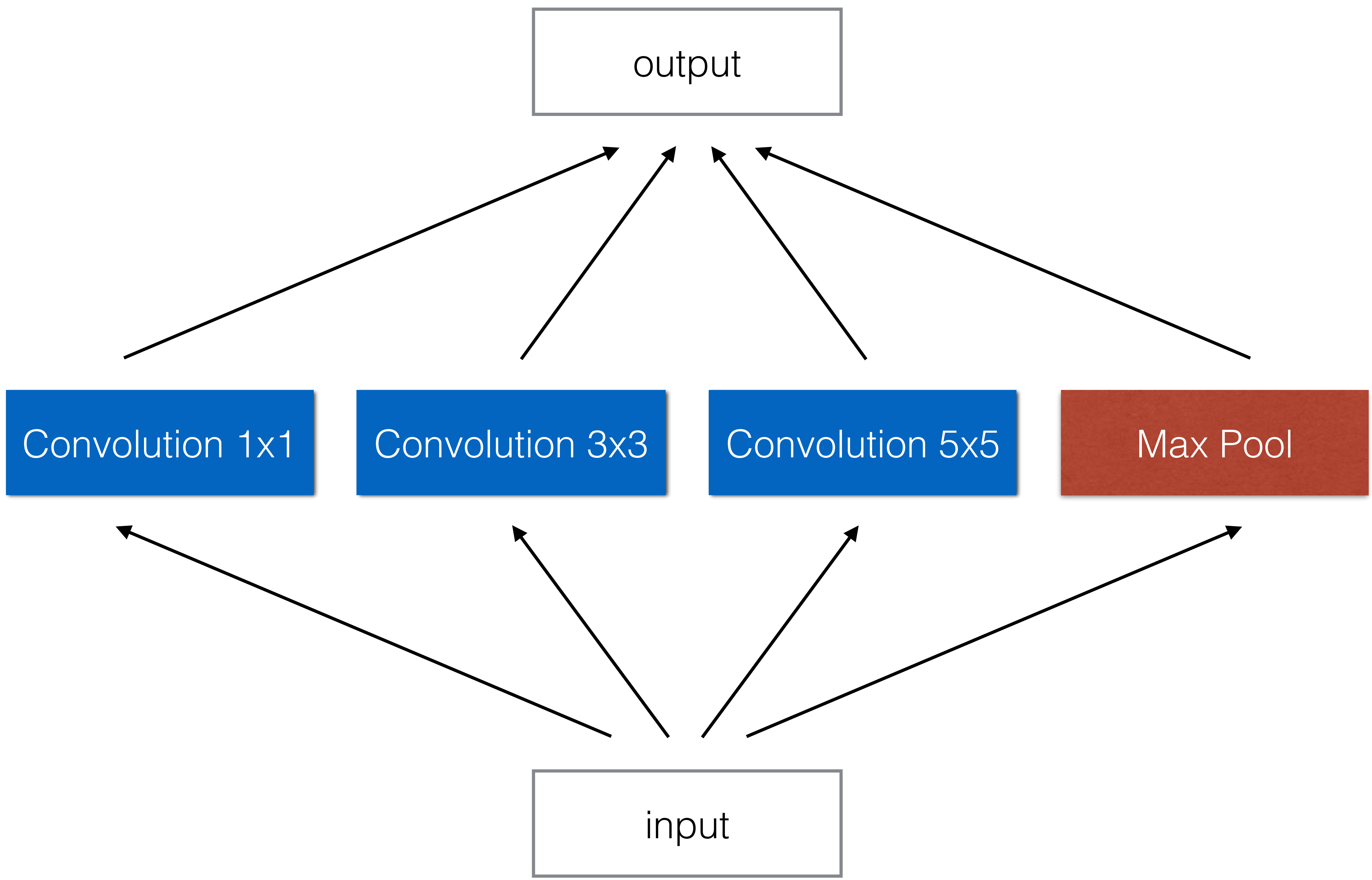
“Network-in-network” constrains representation.

- “*Network-in-network”* architecture demonstrated impressive performance on ImageNet Challenge.



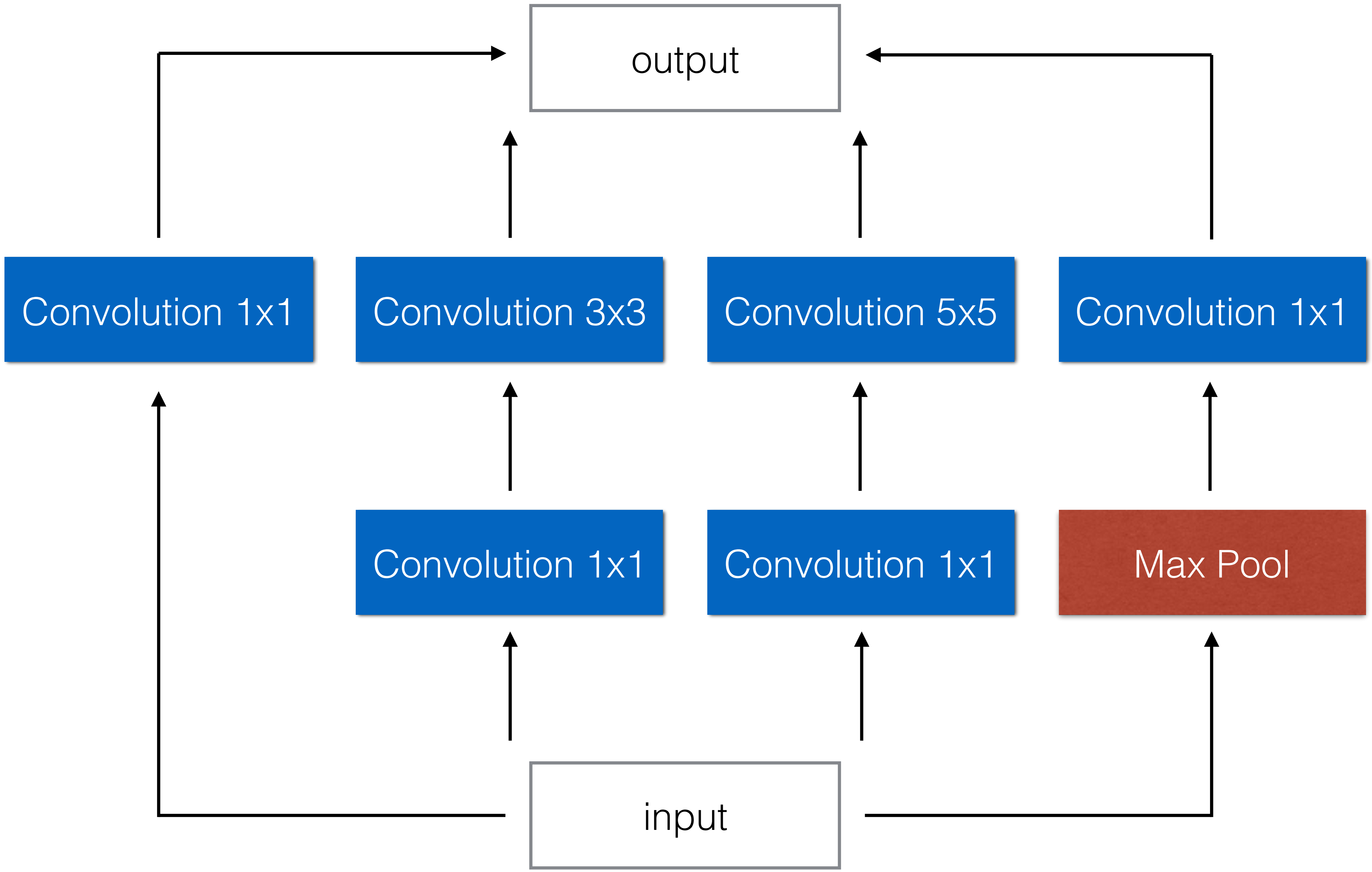
- Restrict the representational power and may reduce the number of matrix multiplications.

Network in network.
M Lin, Q Chen, and S Yan (2013)



Going Deeper with Convolutions
C Szegedy et al (2014)

Employ multi-scale and dimensional reduction.



Going Deeper with Convolutions
C Szegedy et al (2014)

Summary of Inception architecture.

- Multi-scale architecture to mirror correlation structure in images.
- Dimensional reduction to constrain representation along each spatial scale.

Going Deeper with Convolutions
C Szegedy et al (2014)

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- Beyond image recognition

Dist-Belief
Inception

batch normalization
adversarial training

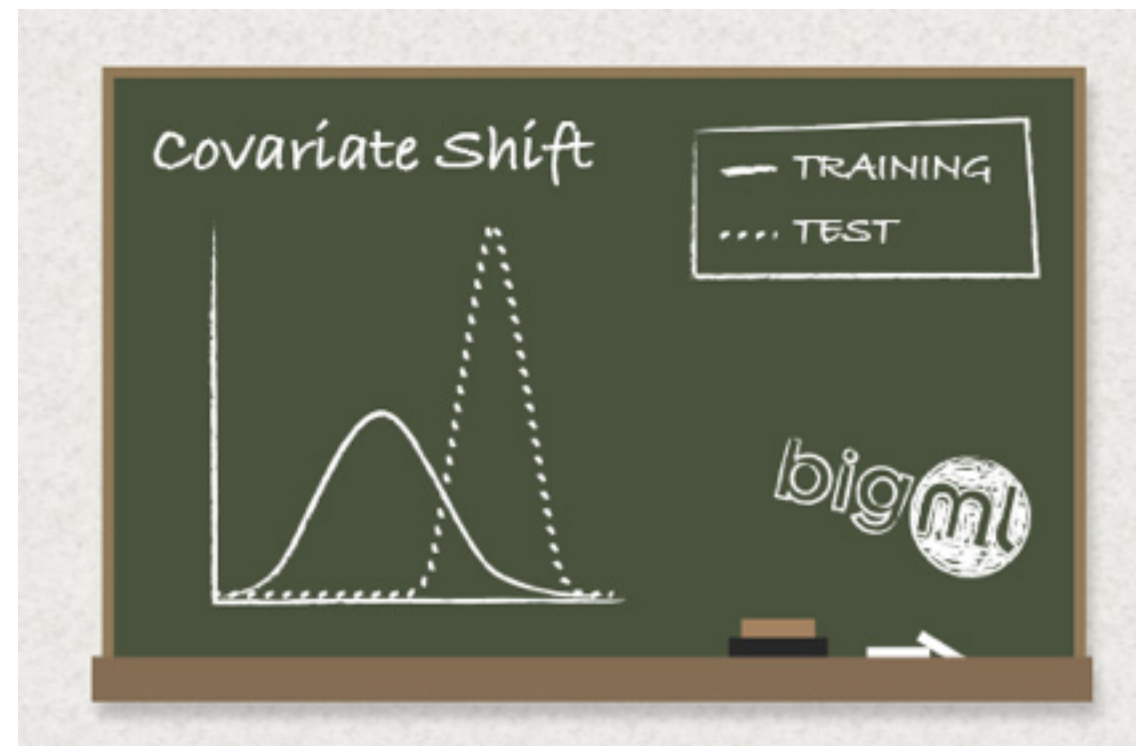
DeViSE
Show-And-Tell

DRAW
video

Covariate shifts are problematic in machine learning

- Traditional machine learning must contend with *covariate shift* between data sets.

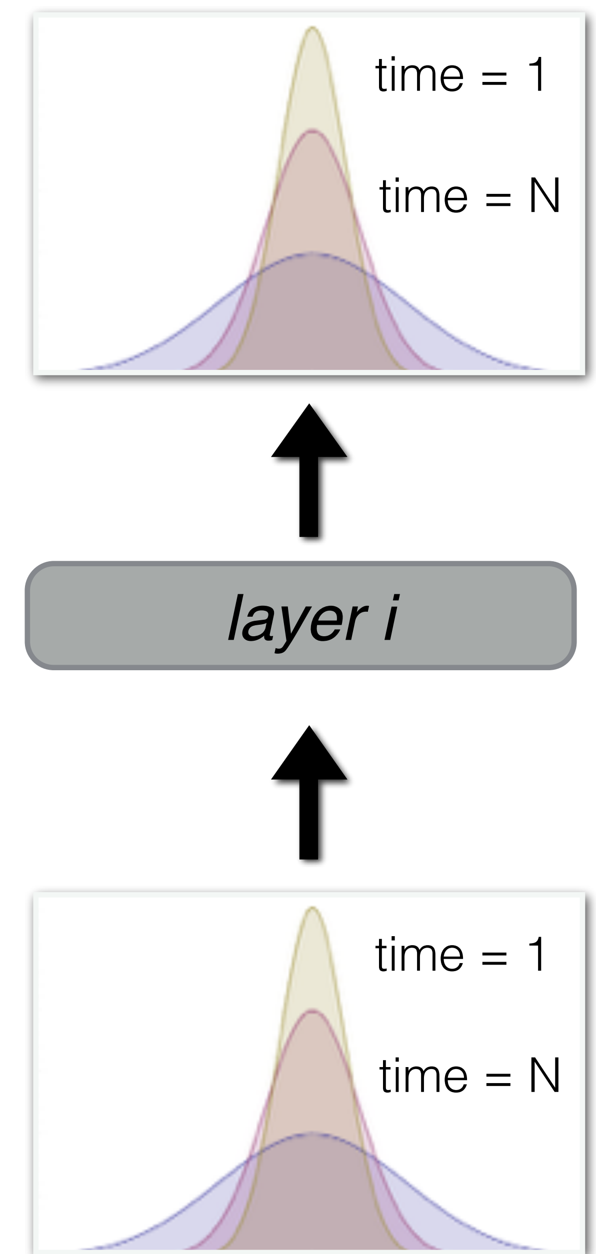
- Covariate shifts must be mitigates through *domain adaptation*.



blog.bigml.com

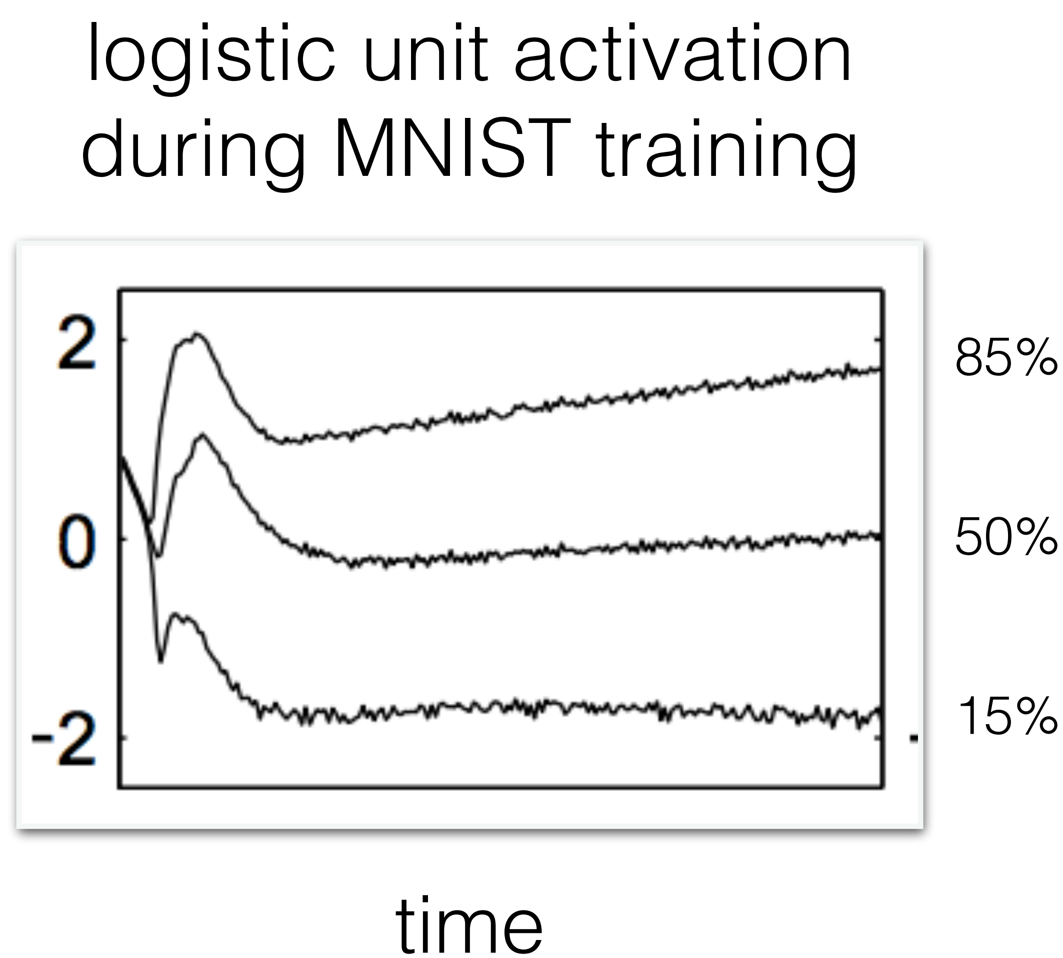
Covariate shifts are problematic in machine learning

- Traditional machine learning must contend with *covariate shift* between data sets.
- Covariate shifts must be mitigates through *domain adaptation*.



Covariate shifts occur between network layers.

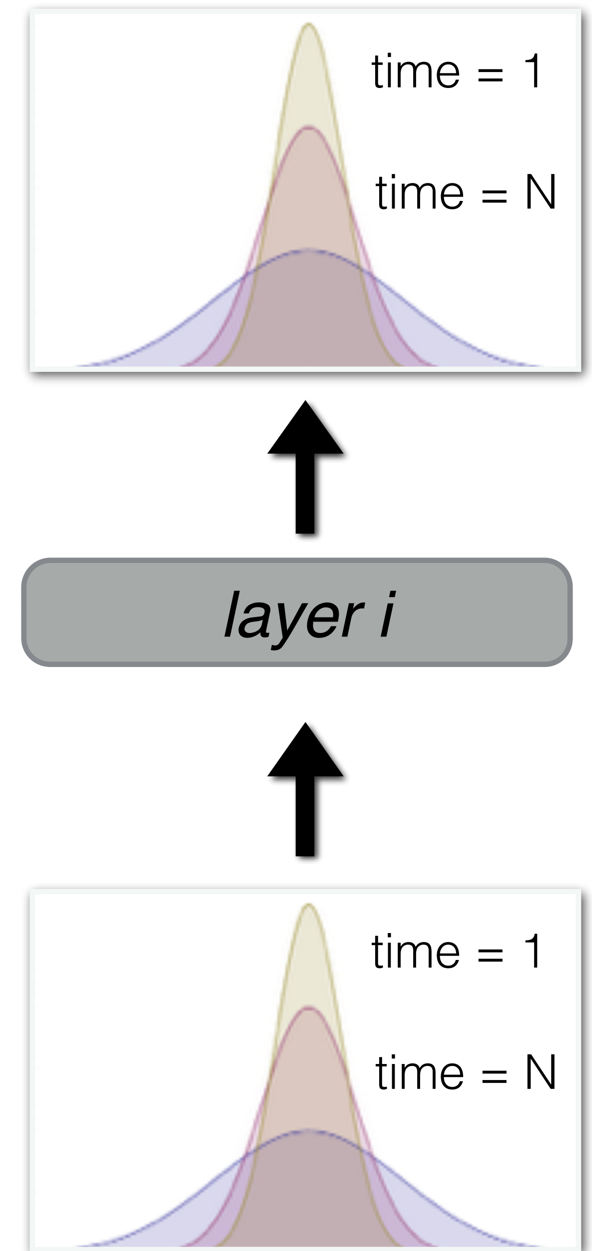
- Covariate shifts occur across layers in a deep network.
- Performing domain adaptation or whitening is impractical in an online setting.



Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
S Ioffe and C Szegedy (2015)

Previous method for addressing covariate shifts

- whitening input data
- building invariances through normalization
- regularizing the network (e.g. dropout, maxout)



I Goodfellow et al (2013)
N Srivastava et al. (2014)

Mitigate covariate shift via batch normalization.

- Normalize the activations in each layer within a mini-batch.

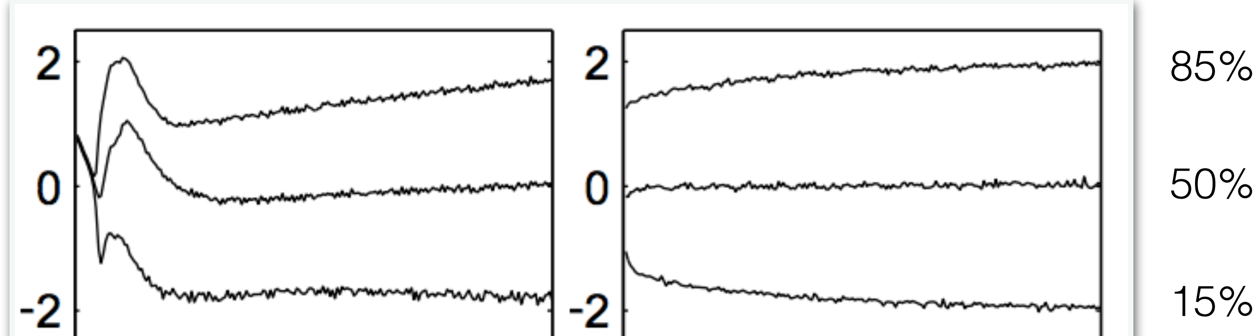
- Learn the mean and variance (γ, β) of each layer as parameters

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$



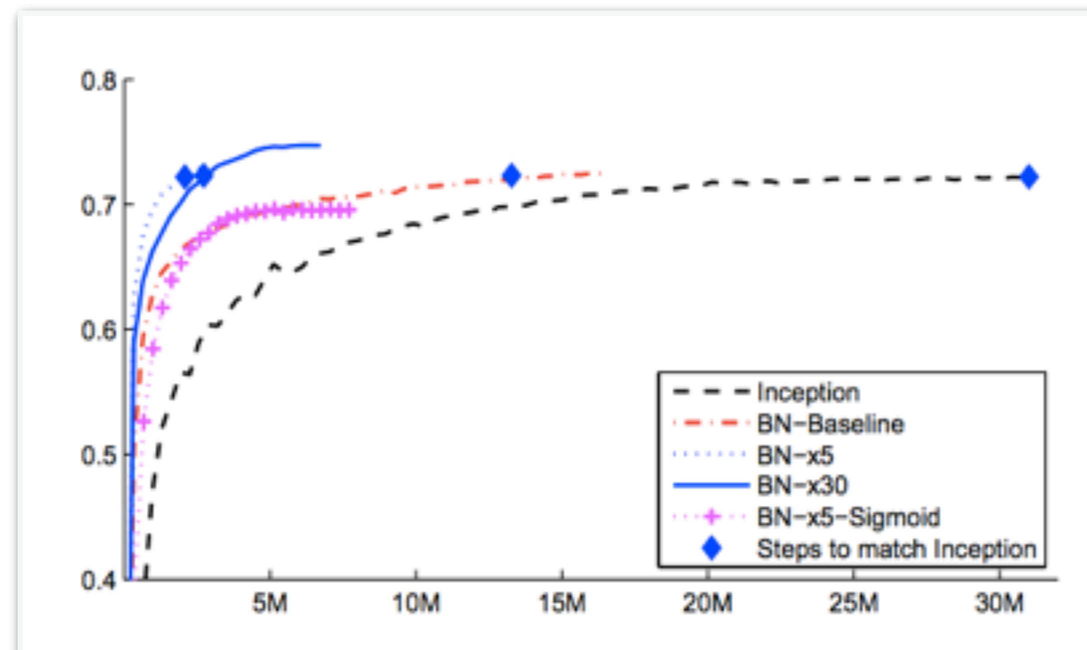
(b) Without BN (c) With BN

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
S Ioffe and C Szegedy (2015)

Batch normalization improves Inception network.

- Multi-layer CNN's train faster with fewer data samples (15x).
- Employ faster learning rates and less network regularizations.
- Achieves state of the art results on ImageNet.

precision @ 1



number of mini-batches

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
S Ioffe and C Szegedy (2015)

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- New directions.

Dist-Belief
Inception

batch normalization
adversarial training

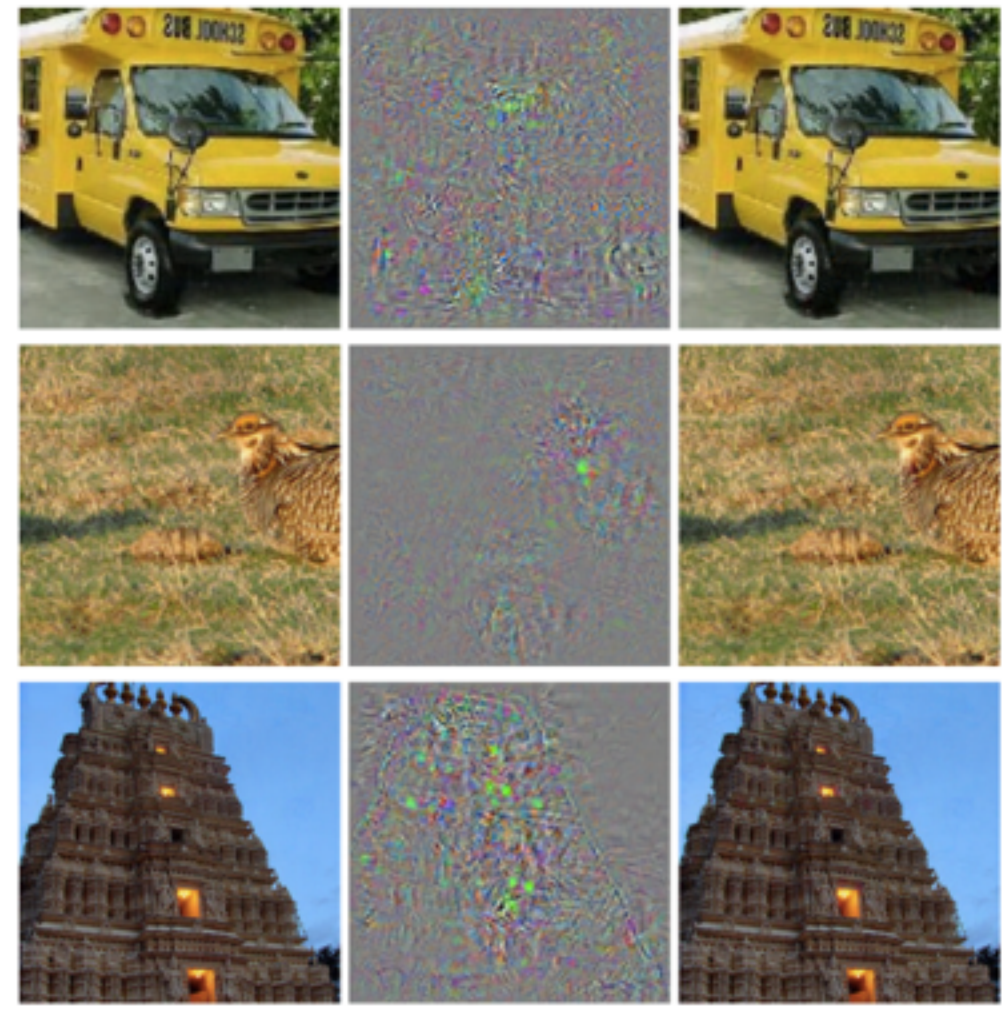
DeViSE
Show-And-Tell

DRAW
video

Machine learning systems can easily be fooled.

- Employ second-order method to search for minimal distortion to create a false classification.

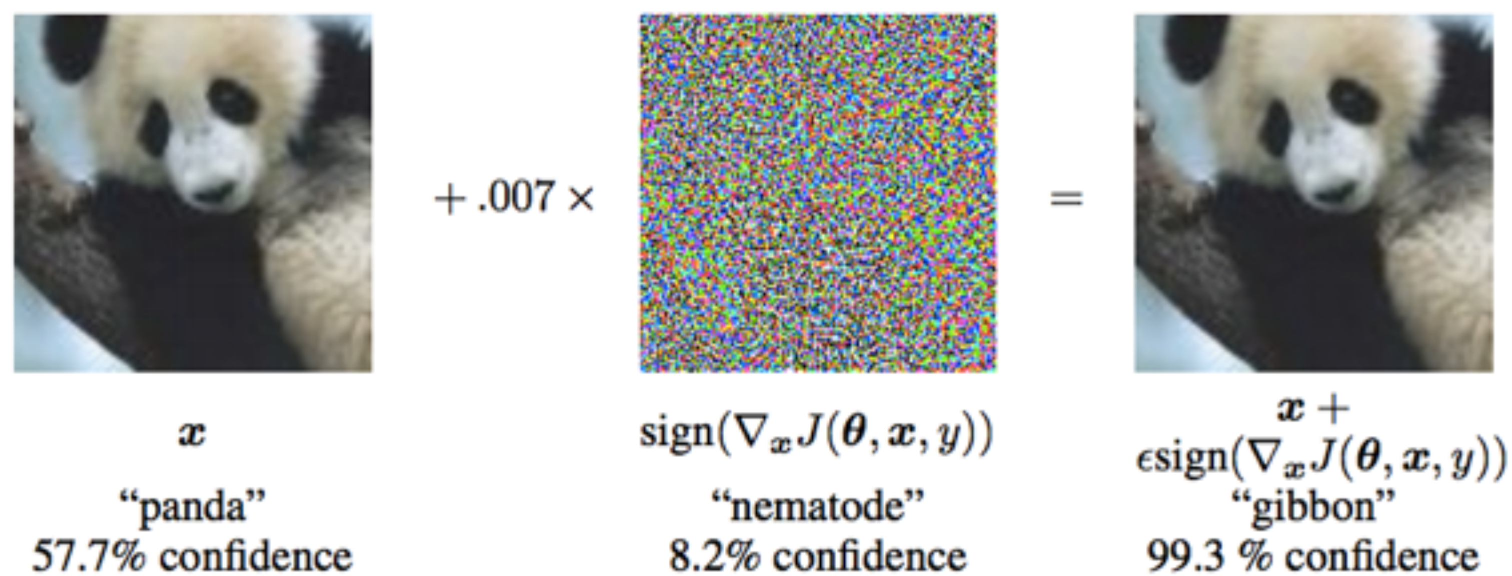
- Generate slight deviations in images that effect almost any image classifier system.



original distortion adversarial

Intriguing Properties of Neural Networks
C Szegedy et al (2013)

Compute adversaries cheaply with gradient.



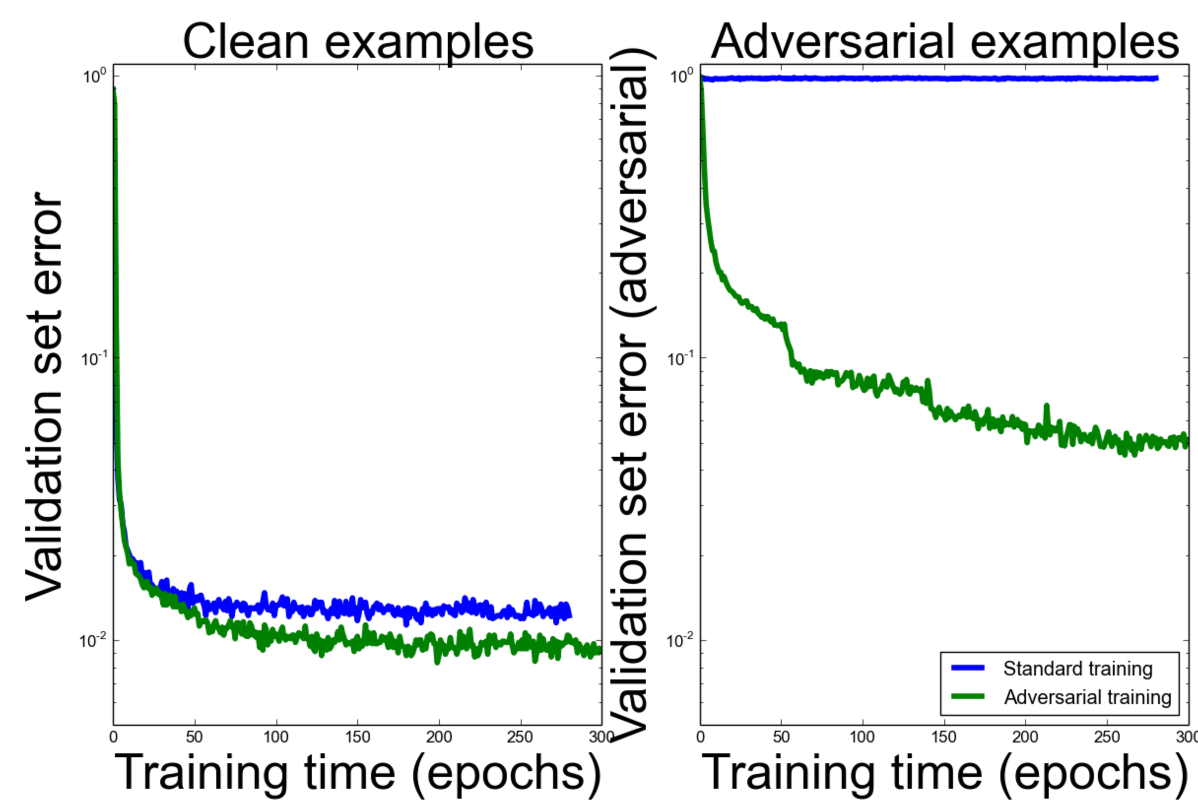
- Generate adversarial examples by back-propagating the loss from the classifier.

- Requires two passes of the network for every image example.

Explaining and Harnessing Adversarial Examples
I Goodfellow et al (2015)

Harnessing adversaries for improves network training.

- Consider adversarial examples as another form of data augmentation.
- Achieved state of the art results on MNIST digit classification (error rate = 0.78%)
- Model becomes resistant to adversarial examples (error rate 89.4% —> 17.9%).



Explaining and Harnessing Adversarial Examples
I Goodfellow et al (2015)

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- Beyond image recognition

Dist-Belief
Inception

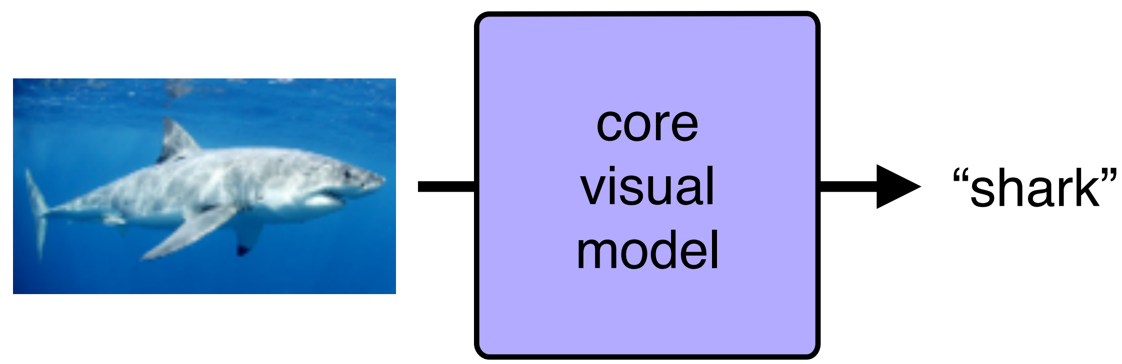
batch normalization
adversarial training

DeViSE
Show-And-Tell

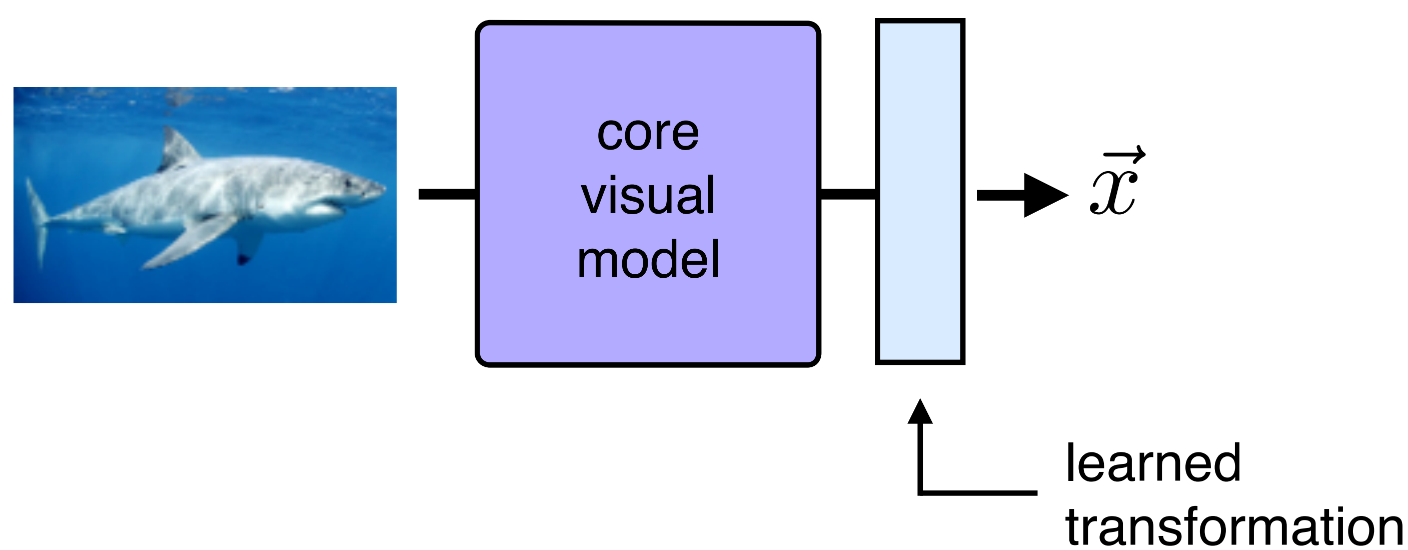
DRAW
video

Classification versus embedding.

- Traditional image models make predictions within a fixed, discrete dictionary.

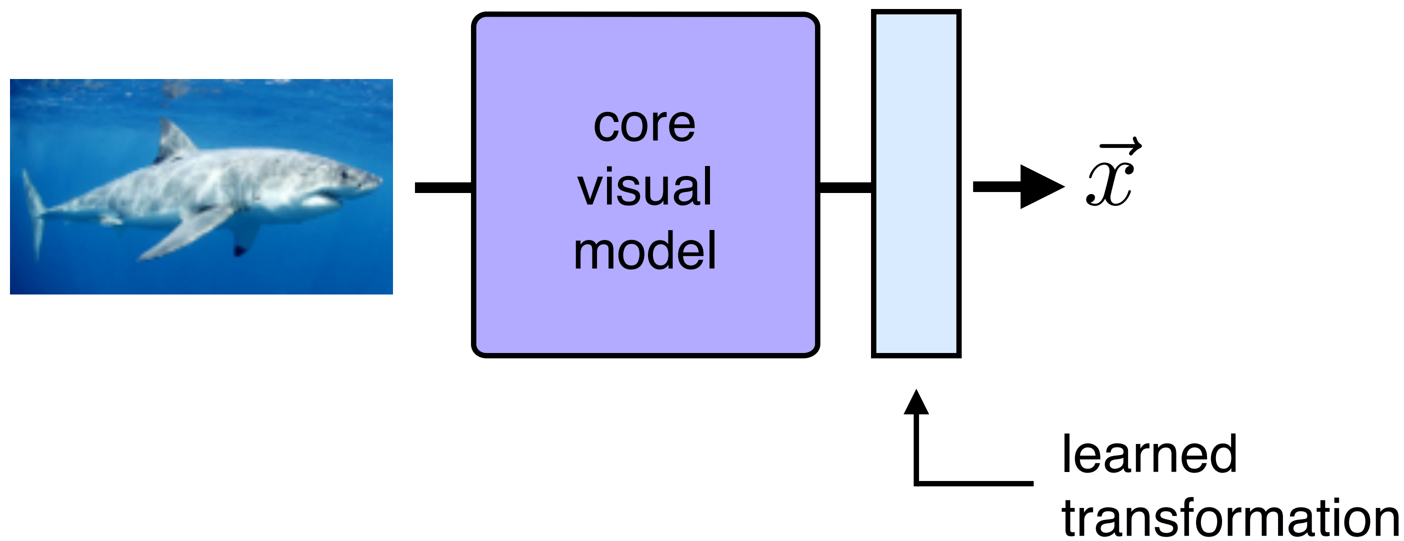


- Why restrict ourselves to classification? Embeddings are far more rich and generic.



Domain transfer in visual domain.

- Embeddings from visual models can be applied “out of the box” to other visual problems.



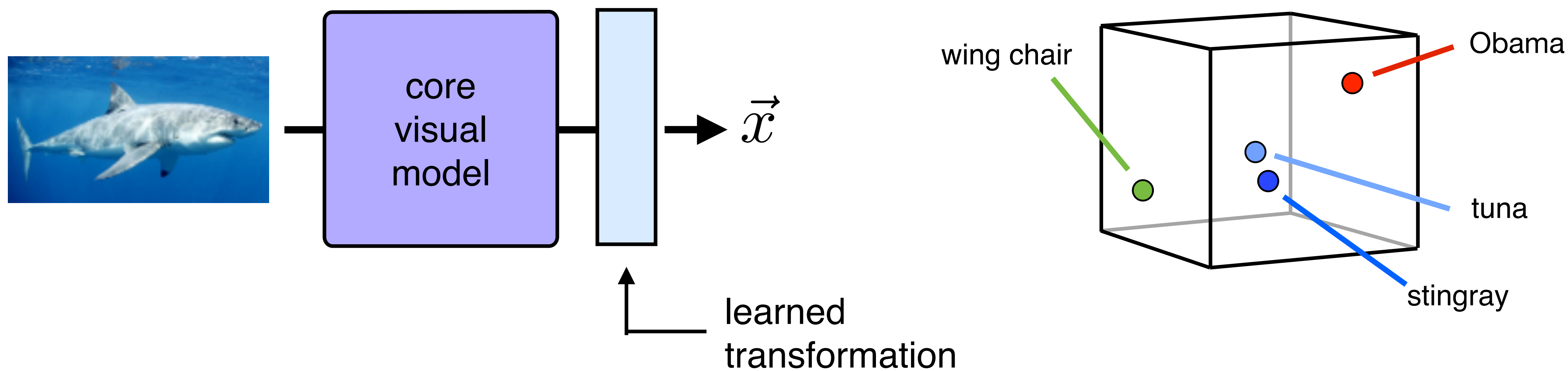
- Embedding are just vectors. Why restrict ourselves to one domain?

DECAF: A deep convolutional activation feature for generic visual recognition
T Darrell et al (2013)

OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks
P Sermanet et al (2014)

Synthesizing vision and language models.

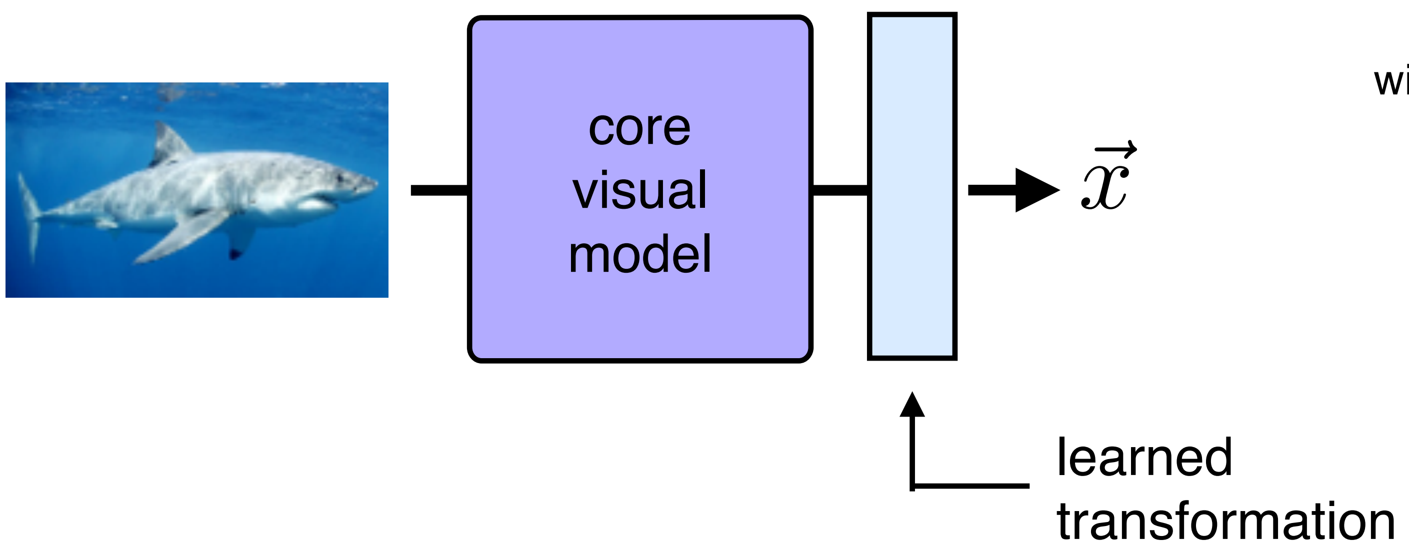
- Train embeddings to predict into language model space.

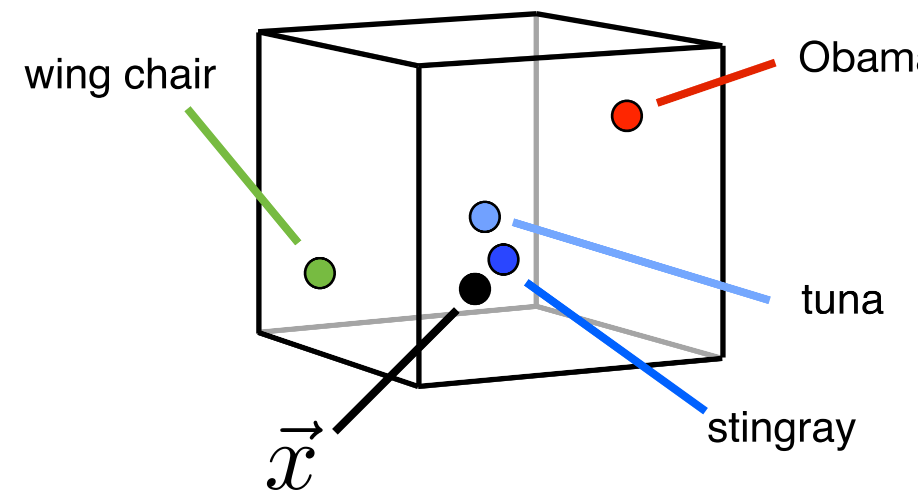


Distributed Representations of Words and Phrases and their Compositionality
T Mikolov et al (2013)
Zero-Shot Learning Through Cross-Modal Transfer
R Socher et al (2013)
DeViSE: A Deep Visual-Semantic Embedding Model
A Frome et al (2013)

Synthesizing vision and language models.

- Train embeddings to predict into language model space.

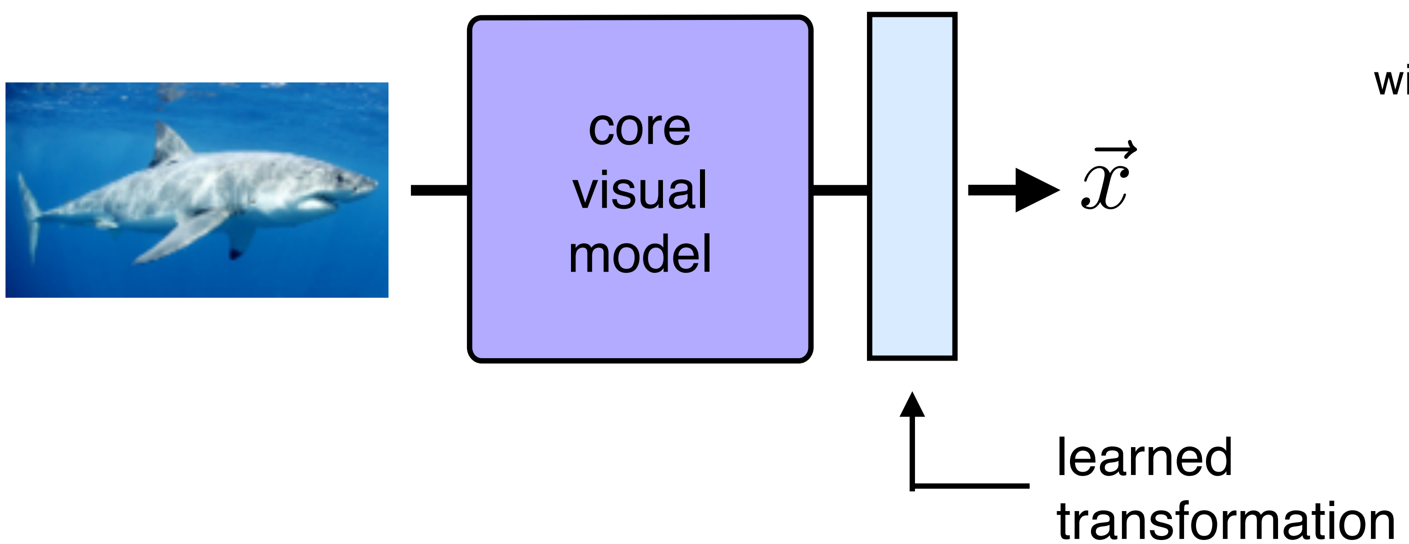


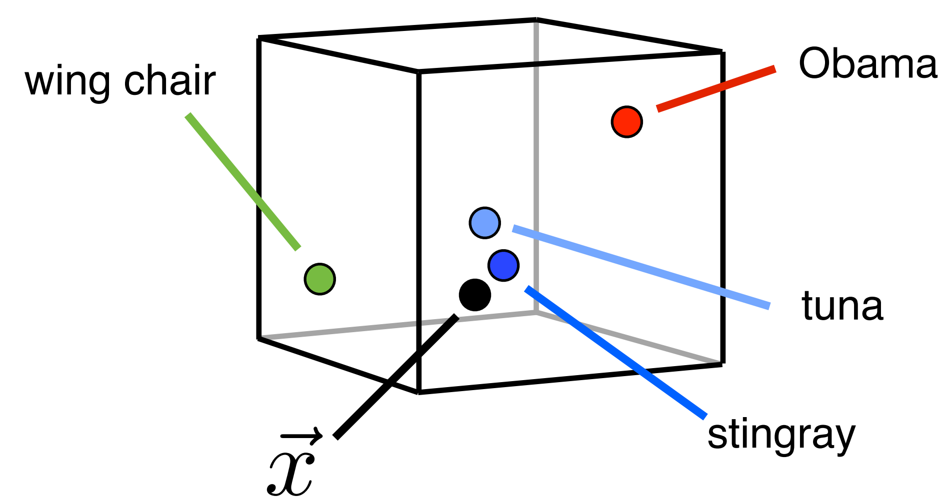


Distributed Representations of Words and Phrases and their Compositionality
T Mikolov et al (2013)
Zero-Shot Learning Through Cross-Modal Transfer
R Socher et al (2013)
DeViSE: A Deep Visual-Semantic Embedding Model
A Frome et al (2013)

Synthesizing vision and language models.

- Train embeddings to predict into language model space.





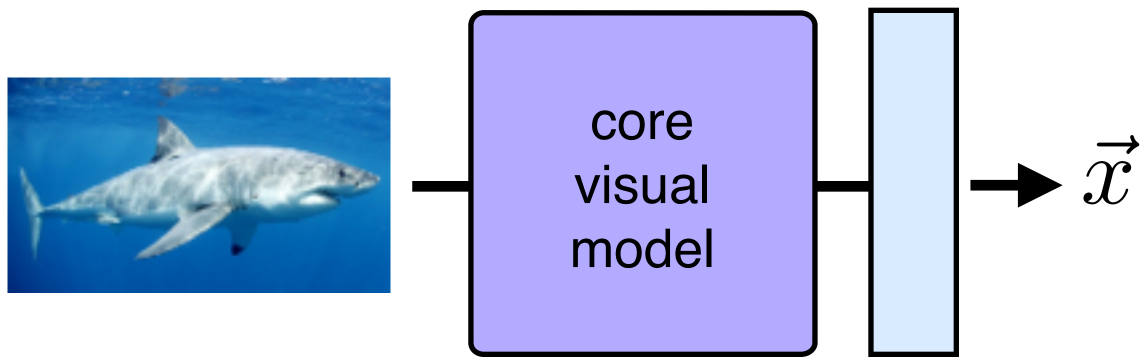
Distributed Representations of Words and Phrases and their Compositionality
T Mikolov et al (2013)
Zero-Shot Learning Through Cross-Modal Transfer
R Socher et al (2013)
DeViSE: A Deep Visual-Semantic Embedding Model
A Frome et al (2013)

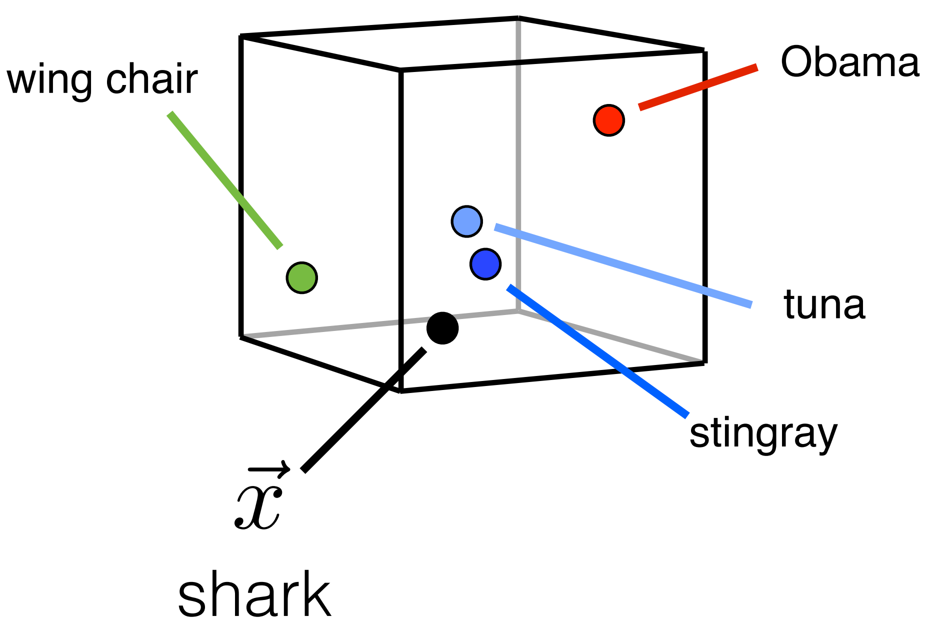
Zero shot learning on unseen image labels.

	"DeViSE", A Frome et al (2013)	A Krizhevsky et al (2012)
	eyepiece, ocular Polaroid compound lens **telephoto lens, zoom lens** rangefinder, range finder	typewriter keyboard tape player reflex camera CD player space bar
	oboe, hautboy, hautbois bassoon **English horn, cor anglais** hook and eye hand	reel punching bag, punch bag, ... whistle bassoon letter opener, paper knife, ...
	barbet patas, hussar monkey, ... **babbler, cackler** titmouse, tit bowerbird, catbird	patas, hussar monkey, ... proboscis monkey, Nasalis ... macaque titi, titi monkey guenon, guenon monkey

Synthesizing vision and language models.

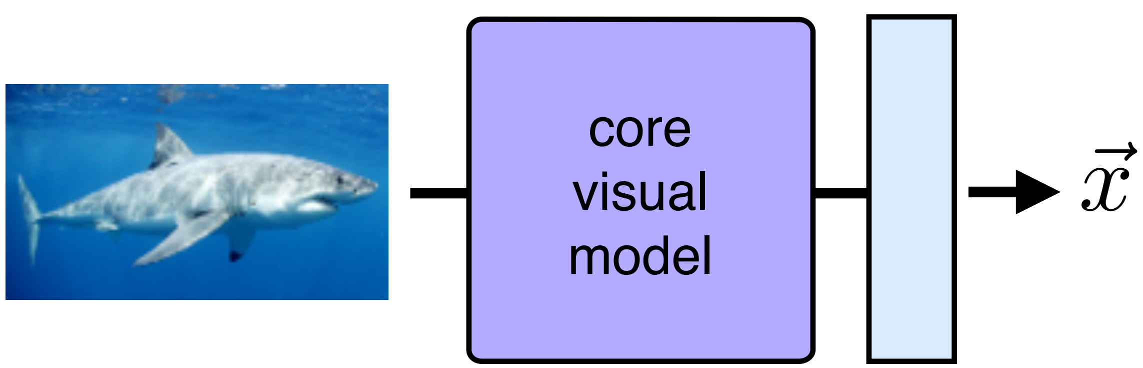
- Language is not just a bag of words but a sequence of words expressing an idea.

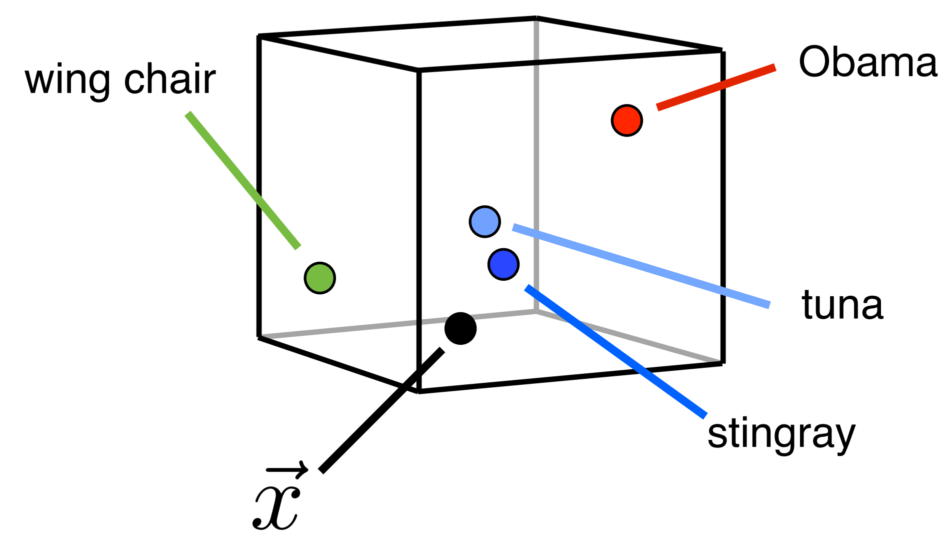




Synthesizing vision and language models.

- Language is not just a bag of words but a sequence of words expressing an idea.

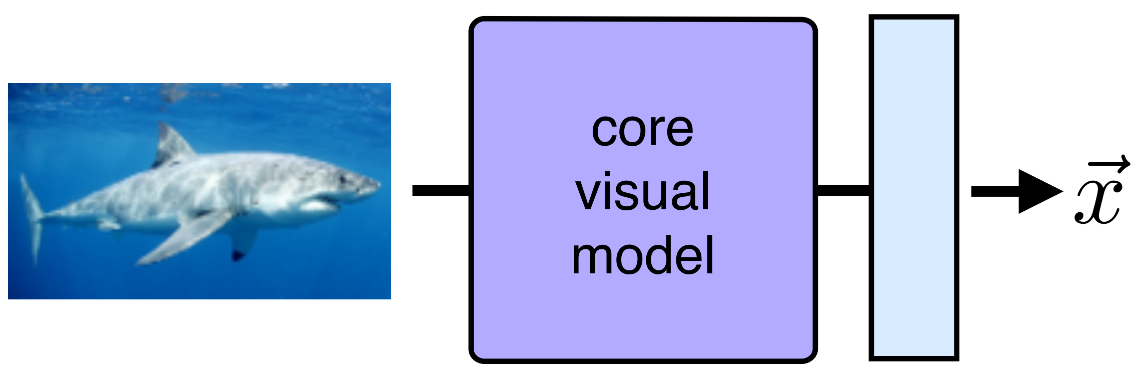




A shark swims in the ocean.

Synthesizing vision and language models.

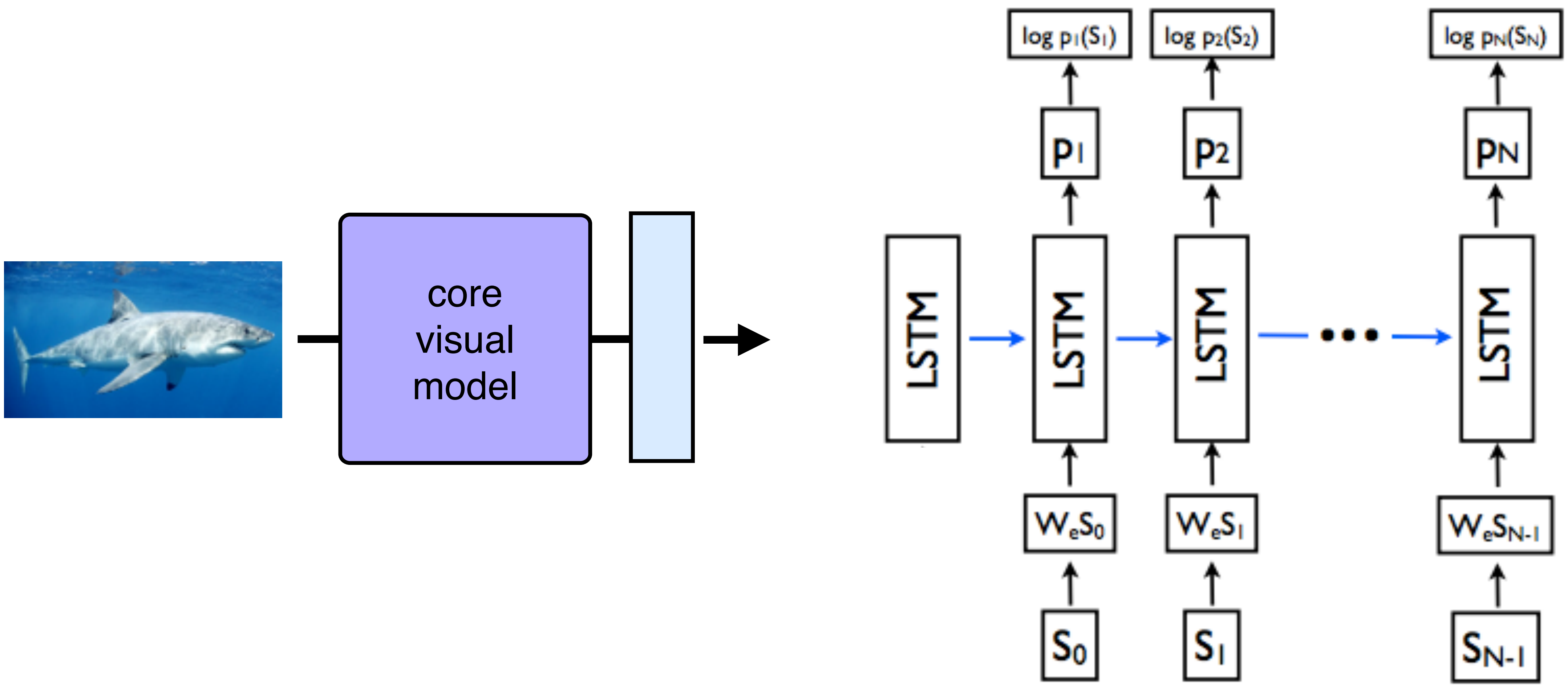
- Language is not just a bag of words but a sequence of words expressing an idea.



A shark swims in the ocean.

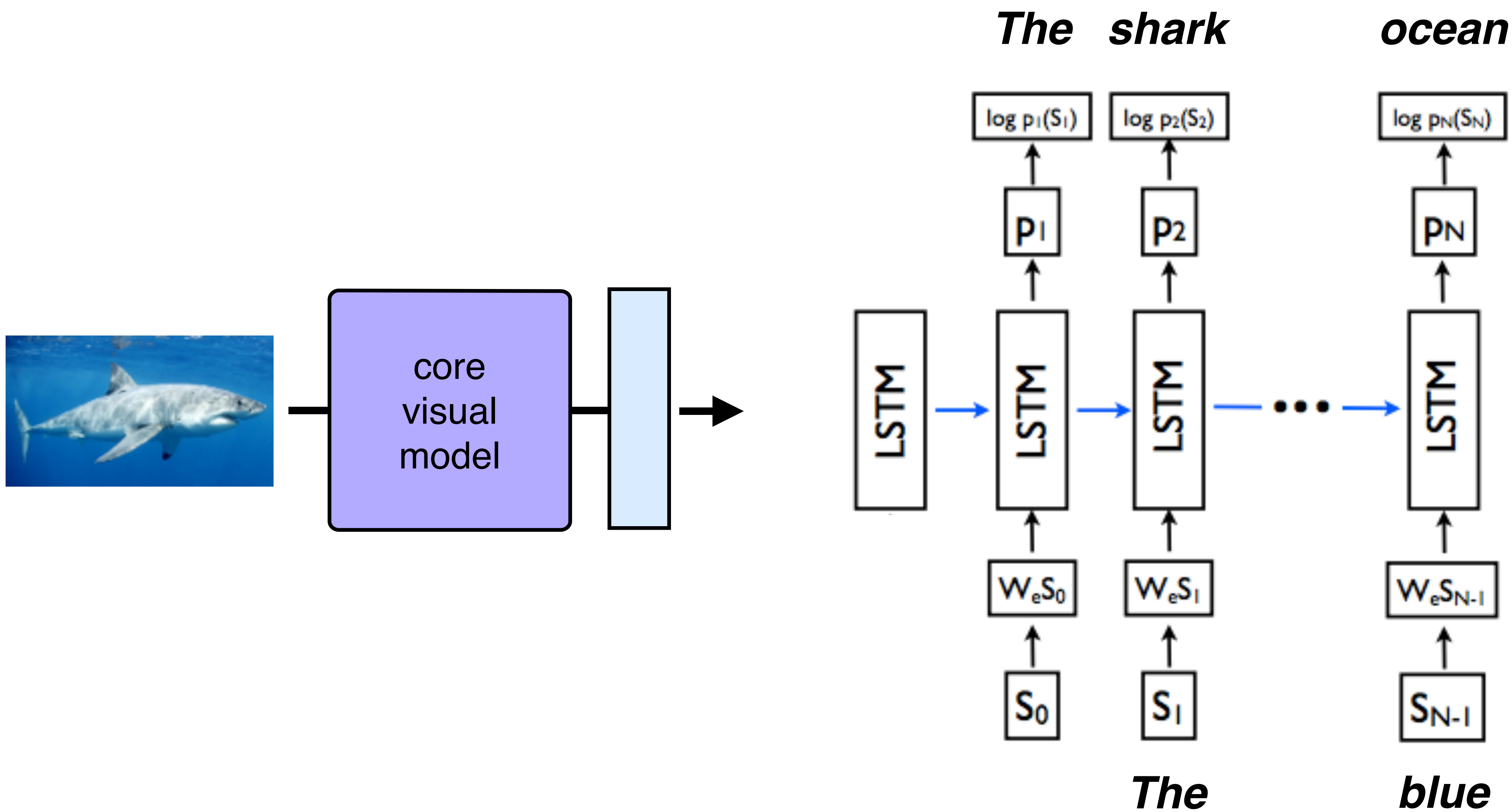
Synthesizing vision and language models.

- Language is not just a bag of words but a sequence of words expressing an idea.



Synthesizing vision and language models.

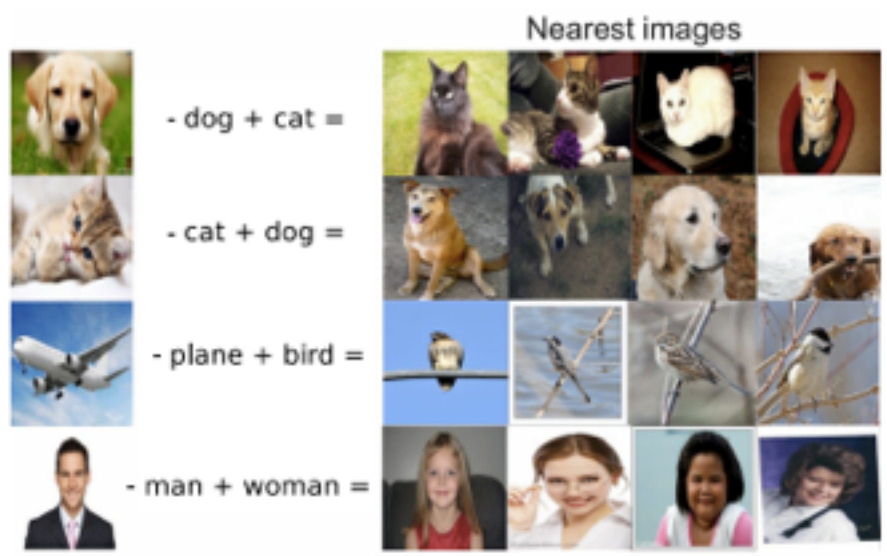
- Language is not just a bag of words but a sequence of words expressing an idea.



Synthesizing vision and language models.

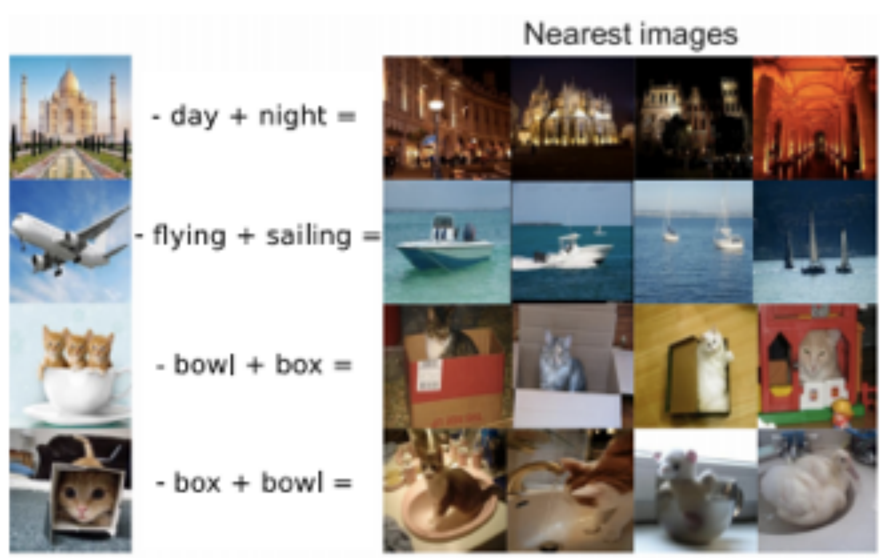
Show and Tell: A Neural Image Caption Generator
O Vinyals et al (2014)

Exploiting the regularities in the language model

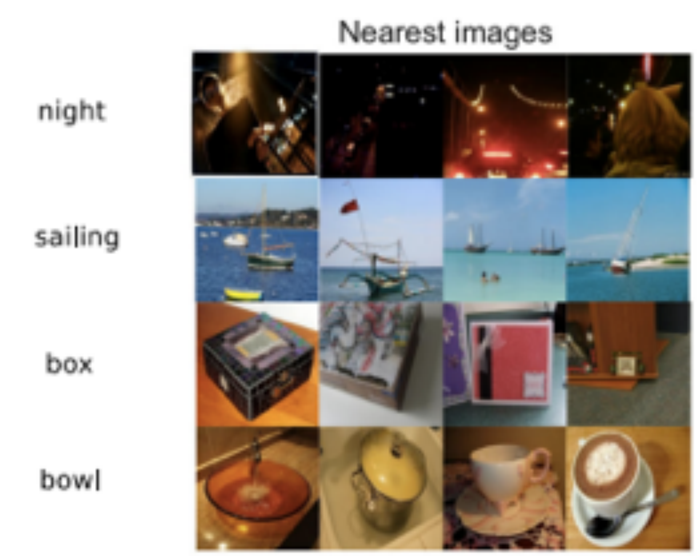


(a) Simple cases

(b) Colors



(c) Image structure

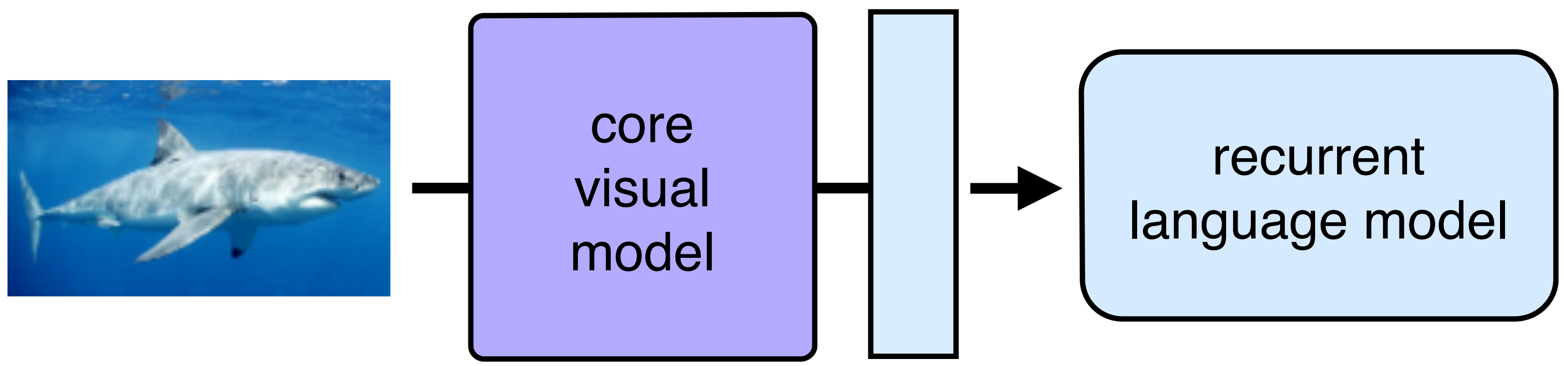


(d) Sanity check

Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
R Kiros, R Salakhutdinov, R Zemel (2014)

Synthesizing vision and language models.

- Language is not just a bag of words but a sequence of words expressing an idea.



Deep Visual-Semantic Alignments for Generating Image Descriptions
A Karpathy and L Fei Fei (2014)
Show and Tell: A Neural Image Caption Generator
O Vinyals et al (2014)
Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
R Kiros, R Salakhutdinov, R Zemel (2014)
Explain Images with Multimodal Recurrent Neural Networks
J Mao, W Xu, Y Yang, J Wang, A Yuille (2014)
Long-term Recurrent Convolutional Networks for Visual Recognition and Description
J Donohue et al (2014)
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
K Xu et al (2015)

The unsung hero is the data.

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin Michael Maire Serge Belongie Lubomir Bourdev Ross Girshick
James Hays Pietro Perona Deva Ramanan C. Lawrence Zitnick Piotr Dollár

a giraffe has it's head up to a small tree.
a giraffe in a pen standing under a tree.
giraffe standing next to a wooden tree-like structure.
a tall giraffe standing next to a tree
a giraffe in an enclosure standing next to a tree.

Outline

- Architectures for building vision models
- New methods for optimization
- Combining vision with language
- Beyond image recognition

Dist-Belief
Inception

batch normalization
adversarial training

DeViSE
Show-And-Tell

DRAW
video

LSTM’s and video

- Consider this a placeholder. Please search for the paper online.

Beyond Short Snippets: Deep Networks for Video Classification
J Ng, M Hausknecht, S Vijayanarasimhan, R Monga, O Vinyals, G Toderici

Naively porting image recognition to video.

- Train a model on ImageNet but score individual video frames from a YouTube video.

fox = 0.27

171.96 sec

fox = 0.63

172.00 sec

fox = 0.45

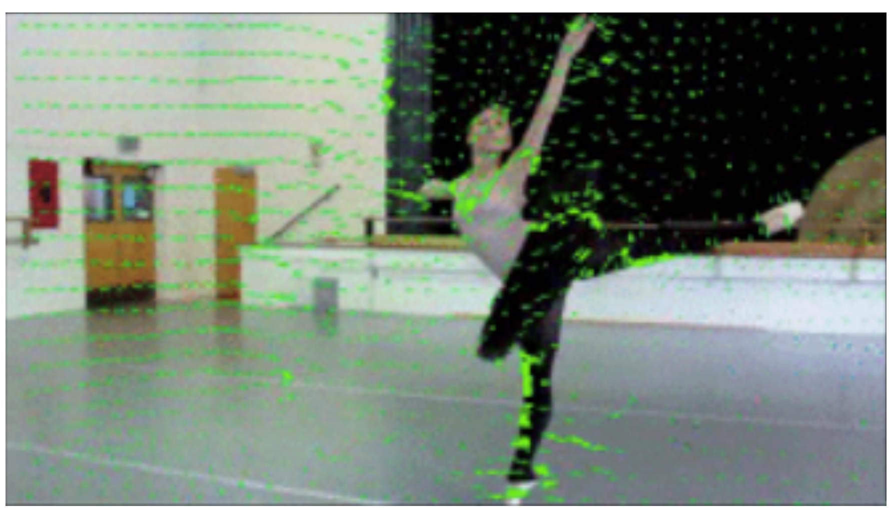
172.03 sec

https://www.youtube.com/watch?v=_AtP7au_Q9w&t=171

Susanna Ricco and Rahul Sukthankar
(unpublished)

Video presents an amazing opportunity.

- Temporal contiguity and motion signals offers an enormous clue for what images should be labeled the same.



tracking →

Susanna Ricco and Rahul Sukthankar
(unpublished)

Outline

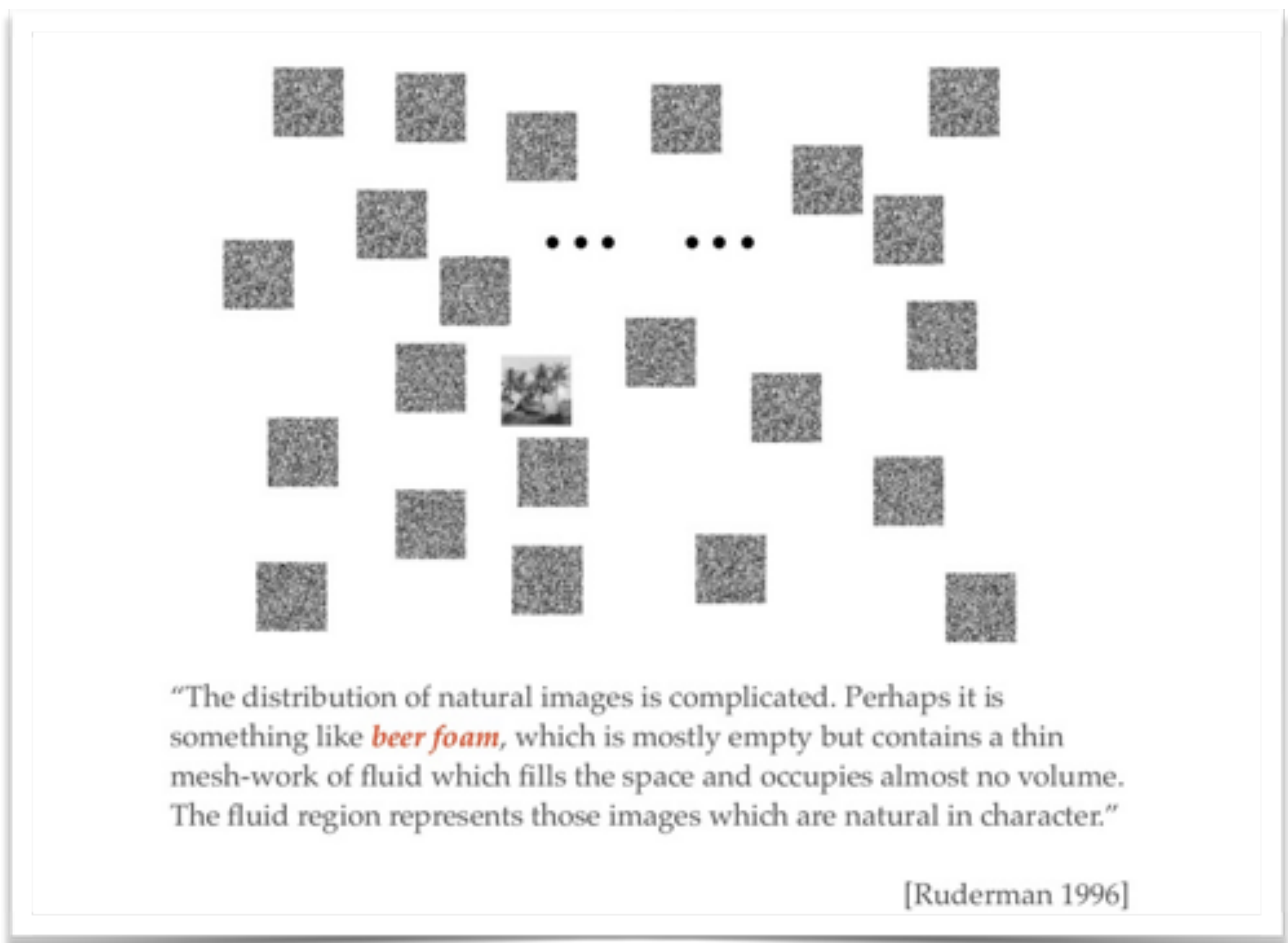
- Architectures for building vision models — Dist-Belief, Inception
- New methods for optimization — batch normalization, adversarial training
- Combining vision with language — DeViSE, Show-And-Tell
- Beyond image recognition — DRAW, video

Synthesizing images is a holy grail.

- Image restoration
  - de-noising, super-resolution, de-mosaicing, in-painting, etc.
- Compression and hashing method
- Debugging and visualizing the state of a CNN network.

Synthesizing images is a challenging domain

- Images reside in a high dimensional space.
- Higher order correlations exist between individual pixels or groups of pixels.



slide from Siwei Lyu

Consider synthesizing an image sequentially.

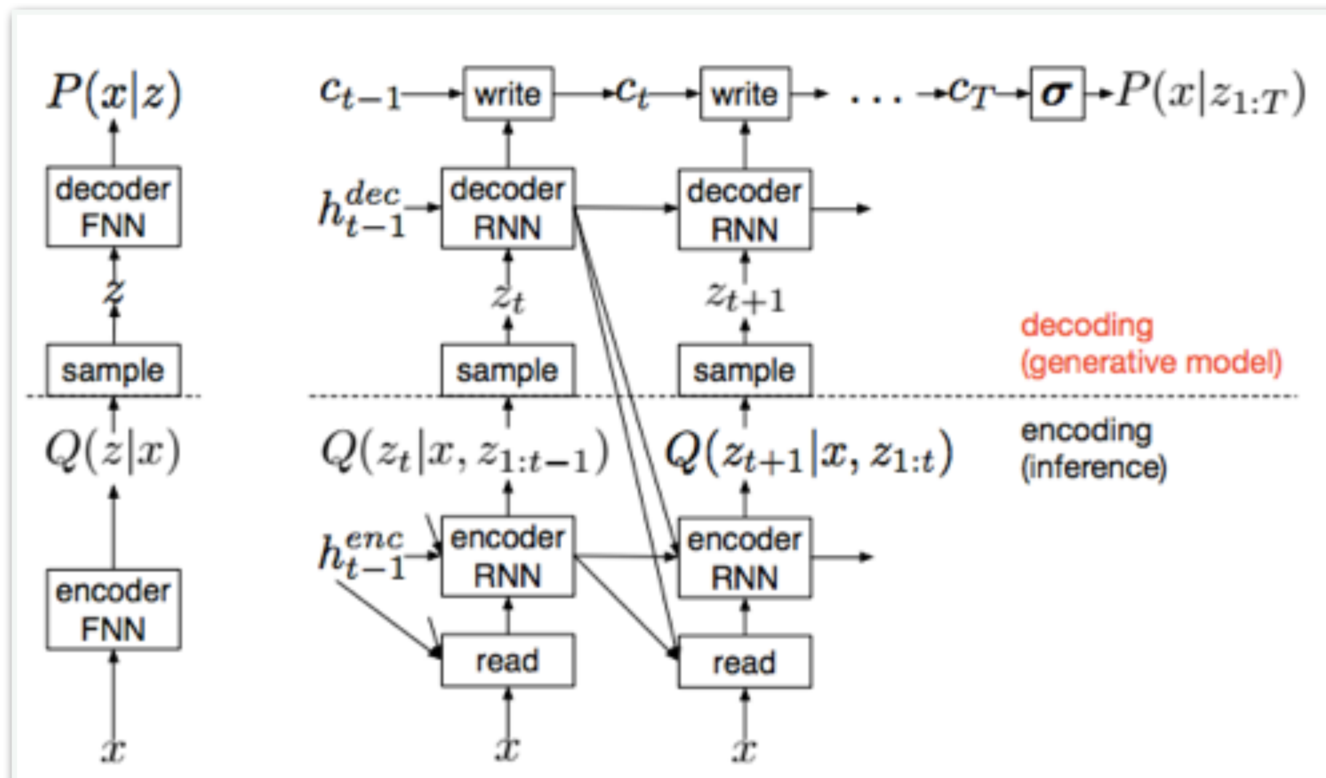
- Network must make a series of consistent predictions.

https://www.youtube.com/watch?v=Zt-7MI9eKEo

DRAW: A Recurrent Neural Network For Image Generation
K Gregor, I Danihelka, A Graves, D Wierstra (2015)

Network employs attention and recurrence.

- Variational auto-encoder + LSTM network.
- Learned selective attention mechanism for drawing and reading an image.

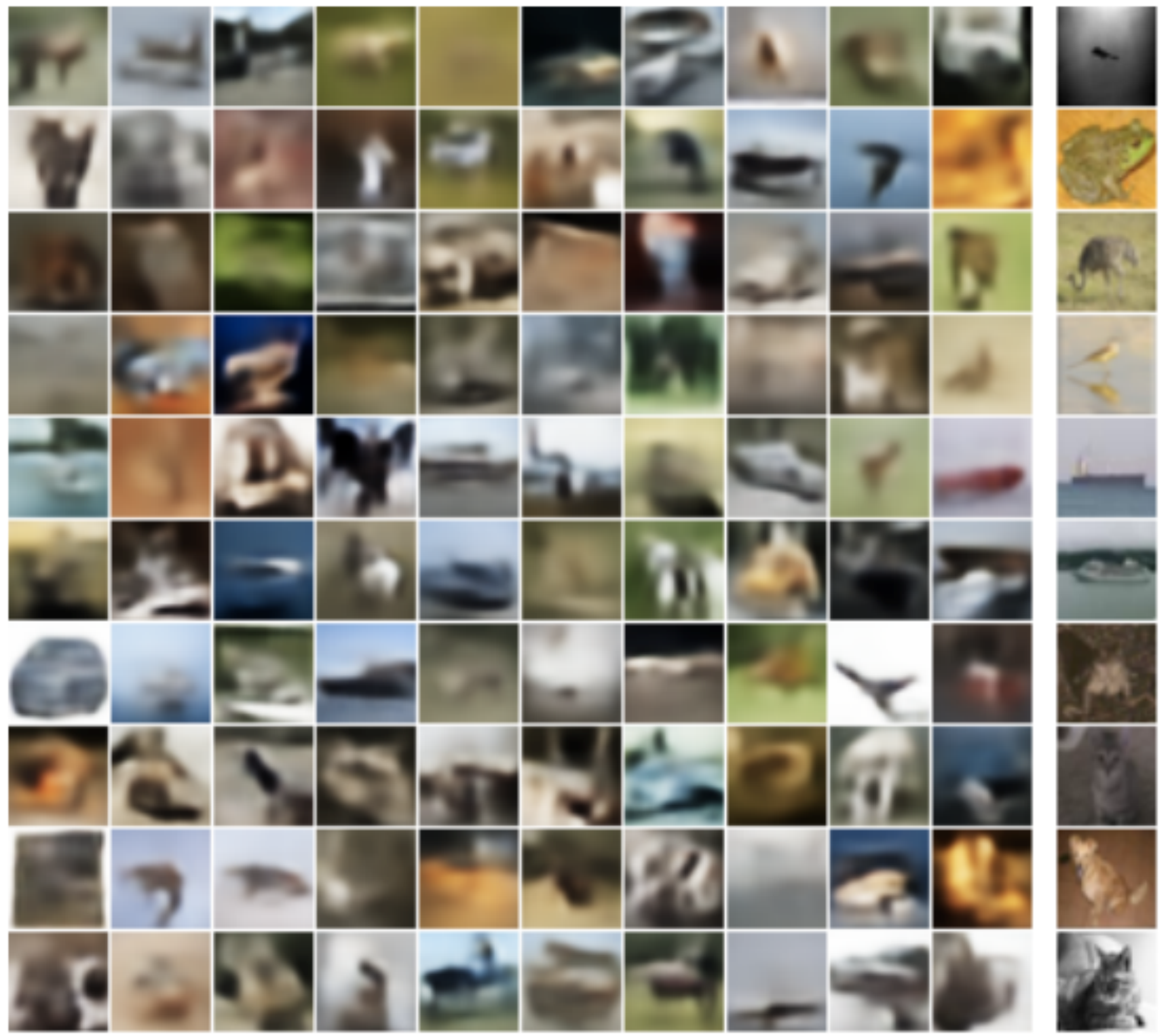


DRAW: A Recurrent Neural Network For Image Generation
K Gregor, I Danihelka, A Graves, D Wierstra (2015)

Synthesized street view house numbers

DRAW: A Recurrent Neural Network For Image Generation
K Gregor, I Danihelka, A Graves, D Wierstra (2015)

Synthesized CIFAR-10 image patches



DRAW: A Recurrent Neural Network For Image Generation
K Gregor, I Danihelka, A Graves, D Wierstra (2015)

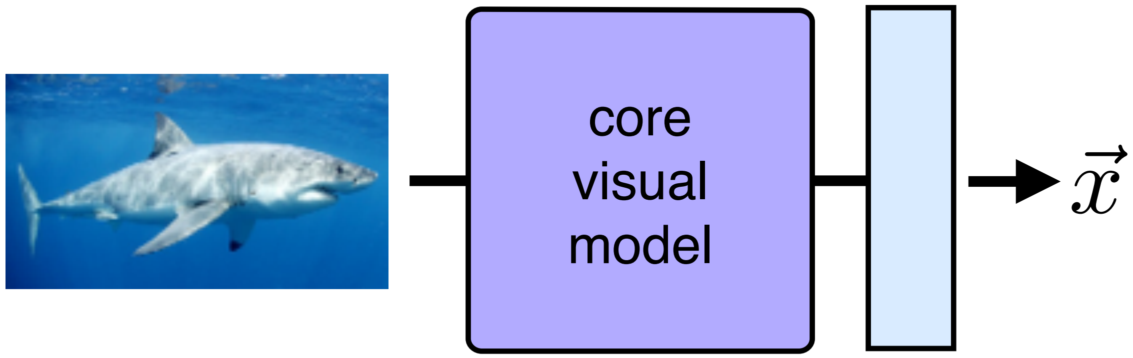
It’s not just about recognizing images.

- Synthesizing images is an open domain to apply convolutional architectures.
- Combining images with other modalities.
- We haven’t even discussed depth.
- How might we curate public data sets to enable this research?

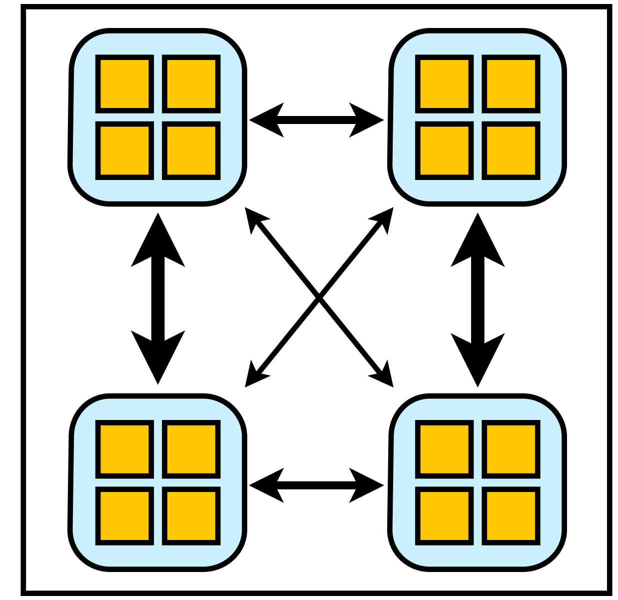
Outline

- Architectures for building vision models — Dist-Belief, Inception
- New methods for optimization — batch normalization, adversarial training
- Combining vision with language — DeViSE, Show-And-Tell
- New directions. — DRAW, video

Vision and Language

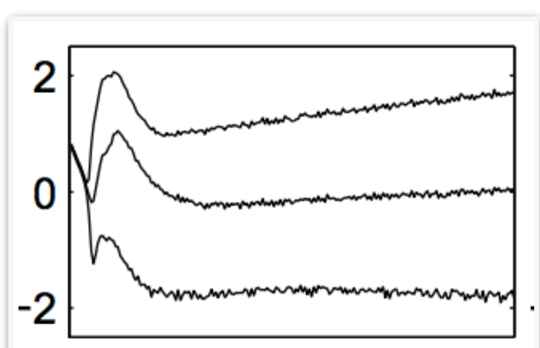


DistBelief



synthesis

Optimization



Inception

Themes

- Vision as a plug-in.
- Transfer learning across modalities.
- Training methods accelerate development of networks