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

<table>
<thead>
<tr>
<th>Technique</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locality constrained linear coding + SVM</td>
<td>NEC &amp; UIUC</td>
</tr>
<tr>
<td>Fisher kernel + SVM</td>
<td>Xerox Research Center Europe</td>
</tr>
<tr>
<td>SIFT features + LI2C</td>
<td>Nanyang Technological Institute</td>
</tr>
<tr>
<td>SIFT features + k-Nearest Neighbors</td>
<td>Laboratoire d'Informatique de Grenoble</td>
</tr>
<tr>
<td>Color features + canonical correlation analysis</td>
<td>National Institute of Informatics, Tokyo</td>
</tr>
</tbody>
</table>

### ImageNet 2011

<table>
<thead>
<tr>
<th>Technique</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed Fisher kernel + SVM</td>
<td>Xerox Research Center Europe</td>
</tr>
<tr>
<td>SIFT bag-of-words + VQ + SVM</td>
<td>University of Amsterdam &amp; University of Trento</td>
</tr>
<tr>
<td>SIFT + ?</td>
<td>ISI Lab, Tokyo University</td>
</tr>
</tbody>
</table>

### ImageNet 2012

<table>
<thead>
<tr>
<th>Technique</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep convolutional neural network</td>
<td>University of Toronto</td>
</tr>
<tr>
<td>Discriminatively trained DPMs</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Fisher-based SIFT features + SVM</td>
<td>ISI Lab, Tokyo University</td>
</tr>
</tbody>
</table>
Convolutional neural networks, revisited

- 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.

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

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

- Winning network contained 60M parameters.
- Achieving scale 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
  - Dist-Belief
  - Inception
- New methods for optimization
  - batch normalization
  - adversarial training
- Combining vision with language
  - DeViSE
  - Show-And-Tell
- Beyond image recognition
  - 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.
One method to achieve scale is parallelization.

Parameter Server

Model Workers

Data Subsets

Large scale distributed deep networks
J Dean et al (2012)
One method to achieve scale is parallelization. Large scale distributed deep networks.

Parameter Server: $p' = p + \Delta p$

Model Workers

Data Shards

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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Top 5 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>Krizhevsky, Suskever and Hinton *</td>
<td>16.4%</td>
</tr>
<tr>
<td>2013</td>
<td>Zeiler and Fergus *</td>
<td>11.5%</td>
</tr>
<tr>
<td>2014</td>
<td>Szegedy et al *</td>
<td>6.6%</td>
</tr>
<tr>
<td>2015</td>
<td>He et al</td>
<td>4.9%</td>
</tr>
<tr>
<td>2015</td>
<td>Ioffe and Szegedy</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

* winner of ImageNet Challenge
Inception is both better and more efficient.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krizhevsky, Suskever and Hinton (2012)</td>
<td>60M</td>
<td>2B</td>
</tr>
<tr>
<td>Zeiler and Fergus (2013)</td>
<td>75M</td>
<td>2B+</td>
</tr>
<tr>
<td>Szegedy et al (2014)</td>
<td>5M</td>
<td>1.5B</td>
</tr>
</tbody>
</table>
Inception is both better and more efficient.

Krizhevsky, Suskever and Hinton (2012)

<table>
<thead>
<tr>
<th>params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>60M</td>
<td>2B</td>
</tr>
</tbody>
</table>

Zeiler and Fergus (2013)

<table>
<thead>
<tr>
<th>params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>75M</td>
<td>2B+</td>
</tr>
</tbody>
</table>

Szegedy et al (2014)

<table>
<thead>
<tr>
<th>params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>5M</td>
<td>1.5B</td>
</tr>
</tbody>
</table>
Inception is both better and more efficient.

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Params</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krizhevsky, Suskever and Hinton (2012)</td>
<td>60M</td>
<td>2B</td>
</tr>
<tr>
<td>Zeiler and Fergus (2013)</td>
<td>75M</td>
<td>2B+</td>
</tr>
<tr>
<td>Szegedy et al (2014)</td>
<td>5M</td>
<td>1.5B</td>
</tr>
</tbody>
</table>

“inception” module
Natural images are locally heavily correlated.
Filter activations reflect image correlations

convolutional filter

image
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.
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.

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.
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
Covariate shifts are problematic in machine learning.

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

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

[blog.bigml.com](http://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 ($\gamma, \beta$) of each layer as parameters.

$$\mu_B \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i \quad \text{// mini-batch mean}$$
$$\sigma^2_B \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 \quad \text{// mini-batch variance}$$
$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma^2_B + \epsilon}} \quad \text{// normalize}$$
$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad \text{// scale and shift}$$

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.

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

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
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
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
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.


eyepiece, ocular
Polaroid
compound lens
**telephoto lens, zoom lens**
rangefinder, range finder

**oboie, hautboy, hautbois**
bassoon
**English horn, cor anglais**
hook and eye
hand

barbet
**patas, hussar monkey, ...**
**babbler, cackler**
titmouse, tit
bowerbird, catbird

typewriter keyboard
tape player
reflex camera
CD player
**space bar**

reel
punching bag, punch bag, ...
whistle
bassoon
letter opener, paper knife, ...

**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.

\[ A \text{ shark swims in the ocean.} \]
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

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.
The unsung hero is the data.
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.

http://www.youtube.com/watch?v=_AtP7au_Q9w&t=171
Video presents an amazing opportunity.

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

“painting of woman”

prediction
Synthesizing images is a challenging domain

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

“The distribution of natural images is complicated. Perhaps it is something like beer foam, which is mostly empty but contains a thin mesh-work of fluid which fills the space and occupies almost no volume. The fluid region represents those images which are natural in character.”

[Ruderman 1996]
Consider synthesizing an image sequentially.

- Network must make a series of consistent predictions.

https://www.youtube.com/watch?v=Zt-7Ml9eKEo
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
Synthesized street view house numbers
Synthesized CIFAR-10 image patches

DRAW: A Recurrent Neural Network For Image Generation
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
- New methods for optimization
- Combining vision with language
- New directions.
Vision and Language

DistBelief

Optimization

Inception

core visual model

\( \vec{x} \)

synthesis
Themes

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