Investigating Data Augmentation Strategies for Advancing Deep Learning Training

Winston H. Hsu (徐宏民)
Professor, National Taiwan University
Director, NVIDIA AI Lab (NTU)

Office: R512, CSIE Building
Communication and Multimedia Lab (通訊與多媒體實驗室)
http://winstonhsu.info

March 26, 2018

Outline

- Why data augmentation in deep learning?
- Data augmentation strategies by
  - Data crawling
  - Weakly supervised learning (least effort for data)
  - Data transformation
  - Synthesizing
- Summary
Deep Learning – A Paradigm Shift in Machine Learning

- Competitive “deep” neural network
- Automatic feature learning (convolution)
- Huge improvements in image/video recognition tasks; so do in audio/speech applications; but marginally in text analytics

For example, classification track in ILSVRC (Top 5 Error)

<table>
<thead>
<tr>
<th>Year</th>
<th>Team</th>
<th>Result</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>SuperVision</td>
<td>0.15</td>
<td>1st Place (CNN)</td>
</tr>
<tr>
<td>2012</td>
<td>ISI</td>
<td>0.26</td>
<td>2nd Place (Conventional)</td>
</tr>
<tr>
<td>2015</td>
<td>MSRA</td>
<td>0.036</td>
<td>1st Place (CNN)</td>
</tr>
</tbody>
</table>

over 96% accuracy if 5 guesses are provided

Why Deep Neural Networks So Powerful?

- “End-to-end training” by
  - Huge training data, GPUs, advanced algorithms, etc.
Deficiencies in Convolutional Neural Networks for Industry Products

- Huge training data required, collected manually now
- Proper network structures?
- Multimodal data ignored
- Bulky parameters & computations

Data is Vital across Learning Paradigms – Example via the (Old) Computer Vision Methods

- PASCAL VOC detection challenge provides realistic benchmark of object detection performance

Zhu et al. Do We Need More Training Data or Better Models for Object Detection? BMVC 2012
Data is Vital for Deep Learning

- AI algorithm is biased?
- Story covered in “Facial Recognition Is Accurate, if You’re a White Guy,” The New York Times, Feb. 9, 2018
- Actually, “gender classification” error caused by lacking quality data in certain categories
  - e.g., the darker the skin, the more errors arise
  - More specific training data will help


Where/How to Get Quality Training Data in an Efficient and Effective Way?

\[
X \rightarrow \Theta \rightarrow Y
\]

- \(X\) training data
- \(\Theta\) training label
- \(Y\) “dog”
- backpropagation
Data Crawling

Rich Image/Videos, Comments, Metadata (GPS, Tags, Time, etc.) in Social Media

Why? Sharing for organization and social communication [Ames, et al., CHI’07]
The First AI-Generated Movie Trailer – Learning from Hundreds of (Horror) Trailers

- Data availability dominates learning model(s)

Image to Poetry by Cross-Modality Understanding

- Joint work with Microsoft Research Asia; deployed live in Microsoft chatbot Xiaolce (小冰)
- Learning from the 519 poets (1920~)
- Hierarchical LSTM-like models for ensuring the intra- and inter-sentence coherence
Netizen-Style Commenting by Learning from Fashion Communities – NetiLook (Public) Dataset

- Contributing the first (large-scale) clothing dataset named NetiLook to discover netizen-style comments; 355,205 images from 11,034 users and 5 million associated comments collected from Lookbook.

- Investigating commenting diversity by topic-parameterized neural networks (NSC)

Social Media are Noisy and Biased

- Subjective and inaccurate for social tagging [Chang, 08]

New York Landmark Labels (Flickr)

<table>
<thead>
<tr>
<th>Locations</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brooklyn Br.</td>
<td>0.38</td>
</tr>
<tr>
<td>Chrysler Building</td>
<td>0.65</td>
</tr>
<tr>
<td>Columbia University</td>
<td>0.30</td>
</tr>
<tr>
<td>Empire State Building</td>
<td>0.18</td>
</tr>
</tbody>
</table>

- Bias in many aspects: gender, length, etc. (e.g., NetiLook Dataset)
Data Annotation – Gaming with a Purpose

- **ESP Game**: labeling image as games [von Ahn, SIGCHI’04]
  - Two people see the same image, and type keywords until they match

- **Other variants**:
  - PeekABoom, Google Labeler, and more in www.gwap.com
  - Label Me

Data Annotation – Advanced Approaches

- Information beyond images / videos
  - speech, semantic network, location, hybrid tag/browse and … mind-reading

[Images and diagrams related to IBM Speech Recognition, Yahoo! ZoneTag, and Brain-Computer Interface are also included.]
Data Annotation – Outsourcing Labeling Task

- **Goal** – outsourcing tasks to a distributed group of people
  - to share the annotation efforts
  - to reduce the personal bias
- **Paid crowd-sourcing by Amazon Mechanical Turk**
- **Mechanisms for ensuring quality**
  - Being completely answered in a HIT
  - Consistence for the “duplicated” questions in a HIT
  - Avoiding robots
  - ...


- **Goal** – data-driven, machine learning-driven approaches are cheaper for collecting (predicting) census data (e.g., income, per capita carbon emission, crime rates, etc.) from Google Street View images
- **Dataset** – the largest fine-grained dataset reported to date consisting of over 2600 classes of cars comprised of images from **Street View** and other web sources, classified by car experts and AMT (object)
“Automatically” Acquiring Effective Training Images for Learning Facial Attributes

- Challenges:
  - Noise
  - Visual diversity
  - Geographical diversity

- Goals
  - Effectiveness for general facial attributes \(\rightarrow\) data/feature selection
  - More diversity in training data \(\rightarrow\) enhanced by contexts

Chen et al. Automatic Training Image Acquisition and Effective Feature Selection From Community-Contributed Photos for Facial Attribute Detection. IEEE TMM 2013

Balancing Content and Context from Social Images

\[ g_k = 1 - \frac{B_G(v_k)}{|G|} \]

\(v_k\): votes from pseudo positives (negatives), weighted by \(x_m\)

\(g_k\): relative \(v_k\) in a grid

\[
\min_p \sum_k \left[ (p_k - t_k)^2 - \beta g_k p_k + \gamma \|p_k\|^2 \right]
\]

Textual Relevance
Visual Consistency
Regularization

\(p_k\): annotation quality; selection indicator \([0, 1]\)
Geographical Diversity for Training Facial Recognizers

- Without geo-context
- With geo-context

<table>
<thead>
<tr>
<th></th>
<th>elder</th>
<th>kid</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Error Rate (%)</td>
<td>26.33 (-0.00)</td>
<td>18.66 (-0.67)</td>
<td>24.50 (-3.00)</td>
</tr>
</tbody>
</table>

Least Effort for the Data
Learning from Noisy Labels – Annotation is Very Expensive (1/2)

- A multi-task network that jointly learns to clean noisy annotations and to accurately classify images
- Using the small clean dataset to learn a mapping between noisy and clean annotations (in a residual manner).

Learning from Noisy Labels – Annotation is Very Expensive (2/2)

- Using the clean labels to directly fine-tune a network trained on the noisy labels does not fully leverage the information
- Clean labels are used to reduce the noise in the large dataset before fine-tuning the network using both the clean labels and the full dataset with reduced noise.
- Experiments in Open Images dataset,
  - **Noisy set**: ~9 million images over 6000 unique classes
  - **Small clean set**: ~40k images.
Network in Network (NIN) – Compact Networks with Global Average Pooling

<table>
<thead>
<tr>
<th>Parameter Number</th>
<th>Performance</th>
<th>Time to train (GTX Titan)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>60 Million (230 Megabytes)</td>
<td>40.7% (Top 1)</td>
</tr>
<tr>
<td>NIN</td>
<td>7.5 Million (29 Megabytes)</td>
<td>39.2% (Top 1)</td>
</tr>
</tbody>
</table>

Global Average Pooling – Huge Parameter Saving by Removing FC Layers

- Global average pooling layer produces spatial average of feature maps as confidence of categories
- Correspondence between feature maps and categories preserved; more meaningful and interpretable
- No parameters (compared to fully connected layers) \(\rightarrow\) prevent overfitting
- Robust to spatial translations of input
Side Product for Global Average Pooling – Visualizing Learned Spatial Correspondence

1. airplane, 2. automobile, 3. bird, 4. cat, 5. deer, 6. dog, 7. frog, 8. horse, 9. ship, 10. truck

Class Activation Map for Weakly Supervised Object Localization (1/3)

- Investigating what CNN is looking in image classification
  - Global Average Pooling (GAP): (1) Does not harm classification results, (2) Remarkable localization ability
  - Class Activation Map (CAM)

The CAMs of two classes from ILSVRC. The maps highlight the discriminative image regions used for image classification, the head of the animal for “briard” and the plates in “barbell”.

Table 2. Localization error on the ILSVRC validation set. Backprop refers to using [23] for localization instead of CAM.

<table>
<thead>
<tr>
<th>Method</th>
<th>top-1 val.error</th>
<th>top-5 val. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet-GAP</td>
<td>56.40</td>
<td>43.00</td>
</tr>
<tr>
<td>VGGnet-GAP</td>
<td>57.20</td>
<td>45.14</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>60.09</td>
<td>49.34</td>
</tr>
<tr>
<td>AlexNet*-GAP</td>
<td>63.75</td>
<td>49.53</td>
</tr>
<tr>
<td>AlexNet-GAP</td>
<td>67.19</td>
<td>52.16</td>
</tr>
<tr>
<td>NIN</td>
<td>65.47</td>
<td>54.19</td>
</tr>
<tr>
<td>Backprop on GoogLeNet</td>
<td>61.31</td>
<td>50.55</td>
</tr>
<tr>
<td>Backprop on VGGnet</td>
<td>61.12</td>
<td>51.46</td>
</tr>
<tr>
<td>Backprop on AlexNet</td>
<td>65.17</td>
<td>52.64</td>
</tr>
<tr>
<td>GoogLeNet-GMP</td>
<td>57.78</td>
<td>45.26</td>
</tr>
</tbody>
</table>

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>supervision</th>
<th>top-5 test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet-GAP (heuristics)</td>
<td>weakly</td>
<td>37.1</td>
</tr>
<tr>
<td>GoogLeNet-GAP</td>
<td>weakly</td>
<td>42.9</td>
</tr>
<tr>
<td>Backprop [23]</td>
<td>weakly</td>
<td>46.4</td>
</tr>
<tr>
<td>OverFeat [22]</td>
<td>full</td>
<td>29.9</td>
</tr>
<tr>
<td>AlexNet [25]</td>
<td>full</td>
<td>34.2</td>
</tr>
</tbody>
</table>
Weakly Supervised Object Detection

- Usual object detector is trained by dataset annotated with bounding boxes
  - Collecting those labels can be very costly and labor intensive.
  - For fields like medical imaging, the labels are even more expensive.
  - Image-level annotation is much easier to get

**Weakly Supervised object detection**

- Aim to train the model to localize the object with only image level supervision (only class label, no bounding boxes)

Weakly Supervised Learning for Localizing Thoracic Diseases – Problem Definition

- Bounding box labels for medical images require professionals to generate the training data. It’s rather time-consuming and expensive.

- Goal – train the network to automatically localize the lesions with only image level supervision. (no bounding box info)
NIH Chest X-Ray 8 Dataset

- Training: 108,948, 8 frontal view X-ray images of 32,717 unique patients with the recording containing disease image class labels (noisy)
- 985 human annotated bounding boxes on 880 images by 8 chest pathologies

Baseline Proposal and Results

- A multi-label classifier with pooling layer (LSE) to increase the localize capability of the network (mixture of GAP and GMP).
- Multiply the weights from prediction layers with the conv feature map to generate the activation heatmap of a specific class, similar to CAM
Data Transformation

Recent Data Augmentation Methods

- Summarized by Thoma in arXiv’17

- Further operations
  - Adding noise
  - Elastic deformations
  - Color casting
  - Vignetting
  - Lens distortion

<table>
<thead>
<tr>
<th>Name</th>
<th>Augmentation Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal flip</td>
<td>2</td>
</tr>
<tr>
<td>Vertical flip</td>
<td>2</td>
</tr>
<tr>
<td>Rotation</td>
<td>$\sim 40$ ($\delta = 20$)</td>
</tr>
<tr>
<td>Scaling</td>
<td>$\sim 14$ ($\delta \in [0.7, 1.4]$)</td>
</tr>
<tr>
<td>Crops</td>
<td>$32^2 = 1024$</td>
</tr>
<tr>
<td>Shearing</td>
<td></td>
</tr>
<tr>
<td>GANs</td>
<td></td>
</tr>
<tr>
<td>Brightness</td>
<td>$\sim 20$ ($\delta \in [0.5, 1.5]$)</td>
</tr>
<tr>
<td>Hue</td>
<td>51 ($\delta = 0.1$)</td>
</tr>
<tr>
<td>Saturation</td>
<td>$\sim 20$ ($\delta = 0.5$)</td>
</tr>
<tr>
<td>Contrast</td>
<td>$\sim 20$ ($\delta \in [0.5, 1.5]$)</td>
</tr>
<tr>
<td>Channel shift</td>
<td></td>
</tr>
</tbody>
</table>
Deep Image: Scaling up Image Recognition – Wu et al., arxiv, 2015 (Baidu)

- **Data augmentation**
  - Cropping, shifting, color casting, lens distortion, vignetting, etc.
  - Training on multi-scale images, including high-resolution ones
    - 512x512 vs. 224x224
  - Hardware/software co-design for parallel computation
    - The number of weights is 212.7M
    - Estimated with 1GB for parameters

- **Configurations**
  - 36 server nodes; each with 4 nvidia Tesla K40; FDR InfiniBand (56Gb/s)
  - Data parallelism in convolutional layers and model parallelism in FC layers
  - SGD synchronization: asynchronous updates

- **Impacts**

<table>
<thead>
<tr>
<th>Team</th>
<th>Year</th>
<th>Place</th>
<th>Top-5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>2012</td>
<td>1</td>
<td>16.42%</td>
</tr>
<tr>
<td>ISL</td>
<td>2012</td>
<td>2</td>
<td>26.13%</td>
</tr>
<tr>
<td>VGG</td>
<td>2012</td>
<td>3</td>
<td>26.98%</td>
</tr>
<tr>
<td>Clarifai</td>
<td>2013</td>
<td>1</td>
<td>11.74%</td>
</tr>
<tr>
<td>NUS</td>
<td>2013</td>
<td>2</td>
<td>12.93%</td>
</tr>
<tr>
<td>ZF</td>
<td>2013</td>
<td>3</td>
<td>13.51%</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>2014</td>
<td>1</td>
<td>6.66%</td>
</tr>
<tr>
<td>VGG</td>
<td>2014</td>
<td>2</td>
<td>7.32%</td>
</tr>
<tr>
<td>MSRA</td>
<td>2014</td>
<td>3</td>
<td>8.06%</td>
</tr>
<tr>
<td>Andrew Howard</td>
<td>2014</td>
<td>4</td>
<td>8.11%</td>
</tr>
<tr>
<td>DeepVision</td>
<td>2014</td>
<td>5</td>
<td>9.51%</td>
</tr>
<tr>
<td>Deep Image</td>
<td></td>
<td>-</td>
<td>5.98%</td>
</tr>
</tbody>
</table>

Table 4: Single model comparison.

- GoogLeNet [21] 7.85%
- VGG [20] 8.65%
- Deep Image 7.42%

Figure 4: Validation set accuracy for different numbers of GPUs.
Securing **robustness** in recognizing consumer photos, often suffering from varying lighting conditions.

- **w/o** augmentation
- **w/** augmentation

Random flip, random crop

- Python open source image augmentation library: `imgaug`
  - Using blur, gaussian noise, brightness, hue, contrast and gray scale
“Shrinking Image” for Learning Super-Resolution (or Face Hallucination)

1/16 or 1/64 size of the original (high quality) one for measuring the reconstruction quality (pixel level or perceptual level)

Mostly with encoder-decoder structure

- Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. CVPR 2017
- Dong et al. Accelerating the Super-Resolution Convolutional Neural Network. ECCV 2016

“Simulated Blurred Images” for Learning De-blurring

Simulated blurred data from the original (high quality) one

- Gong et al., From Motion Blur to Motion Flow: a Deep Learning Solution for Removing Heterogeneous Motion Blur. CVPR 2017
- Liang et al., Dual Motion GAN for Future-Flow Embedded Video Prediction. ICCV 2017
Motivations – labelling images for detection is time-consuming.
  – Every object must be marked with a bounding box.

Augmenting the training data with synthetic images rendered from 3D CAD models (e.g., 3dwarehouse)
Learning Object Detector from 3D Models (2/5)

- How variations in low-level cues affect the features by CNN on the object detection (e.g., PASCAL VOC2007 dataset).
  - Object color, texture and context
  - Synthetic image pose
  - 3D Shape

Learning Object Detector from 3D Models (3/5)

- Object color, Texture and Context

the network has learned to be invariant to the color and texture of the object and its background.
Learning Object Detector from 3D Models (4/5) – Experiments in Synthetic Pose

- Adding side view to front view gives a boost.
- Less invariance.

![Image showing side-view, front-view, and intra-view examples with tables below showing mAP values for different categories and datasets.]

Learning Object Detector from 3D Models (5/5) – Training Object Detection with Limited Real Images

- When the number of real training images is limited, 3D models perform better than traditional RCNN over limited training data.

![Graph showing mAP vs. number of positive real images with various models compared.]
Augmented Reality for Data Generation (1/3) – Motivations

- Creating realistic 3D content is challenging and labor-intensive.
- Real-world images at large scale is easy and directly provides real background appearances without complex 3D models
- **Augmented imagery** generalizes better than *synthetic 3D data* or *limited real data*

![Real scene (KITTI) vs Synthetic scene (Virtual KITTI)](image)

![Real scene augmented with synthetic cars (Ours)](image)

Alhaija et al., Augmented Reality Meets Deep Learning for Car Instance Segmentation in Urban Scenes. BMVC 2017

Augmented Reality for Data Generation (2/3) – Augmentation Pipeline

- Given a set of 3D car models, locations *(manual or automatic)* and environment maps
- Rendering high quality cars and overlay them on top of real images.
- The final post-processing step ensures better visual matching between the rendered and real parts of the resulting image.

![Augmentation Pipeline Diagram](image)
Augmented Reality for Data Generation (3/3) – Impacts Measured by Instance (Car) Segmentation

- More data – real or augmented – are both helpful

![Graph showing AP50 for different data sets](image)

- Augmented foreground cars with real backgrounds are effective

![Images of cars with different backgrounds](image)

Alhaiaja et al., Augmented Reality Meets Deep Learning for Car Instance Segmentation in Urban Scenes. BMVC 2017

Face Recognition (Verification or Identification) by Varying (Multi-Tasking) Loss and Datasets

![Diagram of CNN model](image)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Images</th>
<th>#Persons</th>
<th>#Images per person</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASIA</td>
<td>0.49M</td>
<td>10K</td>
<td>50</td>
</tr>
<tr>
<td>VGGFace</td>
<td>2M</td>
<td>2K</td>
<td>1000</td>
</tr>
<tr>
<td>Megaface2</td>
<td>4.8M</td>
<td>672K</td>
<td>7</td>
</tr>
<tr>
<td>MS-celeb</td>
<td>10M</td>
<td>100K</td>
<td>100</td>
</tr>
<tr>
<td>UMDFace</td>
<td>0.37M</td>
<td>8.5K</td>
<td>40</td>
</tr>
</tbody>
</table>
Problem – Each pair is the same person but predicted to different people (red pairs) by the deep methods because of thick glasses

Why? Few faces with glasses in most face datasets but common in Asian

Glass invariance augmentation by overlapping variant glass models

<table>
<thead>
<tr>
<th></th>
<th>w/o glasses aug.</th>
<th>w/ glasses aug.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>99.207%</td>
<td>99.420%</td>
</tr>
<tr>
<td>Error case on glasses</td>
<td>109</td>
<td>40</td>
</tr>
</tbody>
</table>
(Real-Time) Style Transfer by Perceptual Losses

Johnson et al., Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV 2016

CycleGAN for Synthesizing “Realistic” Training Data

- Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ICCV 2017
- Zhang et al., Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network. CVPR 2018
CycleGAN for Synthesizing “Realistic” Training Data

- Purposes for synthesized medical images
  - as an intermedium in cross-modality image registration or learning
  - as supplementary training samples to boost the generalization capability
- Our modified CybleGAN for multimodal recognition for utilizing MRI and CT (here CT → MRI) in Nasopharyngeal Carcinoma (NPC)

Summary – Utilizing Data Efficiently and Effectively

- data crawling
- transformation
- weakly supervised
- synthesizing
Take Home Messages

- Data are vital for learning paradigms but very costly
- Collecting more training data from public datasets
  - Used for multi-task learning or pre-training
- Augment data with
  - Social media
  - Synthesized data: transformed data, 3D, AR, GAN,
  - Work with the noisy data
  - Weakly supervised methods for minimizing human costs
- Data augmentation is vital for industry applications and will emerge as an important technical component
- Privacy! Privacy! Privacy!
Facebook, LinkedIn: “Winston Hsu”