Example
Intelligent Machines
Deep Representation Learning for Multimedia Data Analysis

Dr. Haojin Yang
Technologies
Why Machine Vision so Hard?
Representative Features

- Raw representations
  - Speech: phoneme
  - Language: letter
  - Image: pixel

\[3^{361} \text{ states} > \text{sum of the universe's atoms}\]

\[256^{3 \times 640 \times 480} \text{ states by using pixel representation}\]
Representative Features
Representative Features
How Kids Know This World
Why has Deep Learning Been so Successful Lately?

- **Large-scale annotated data sets** (e.g., ImageNet: 14 million images in 22k categories; YouTube-8M)
- Deep learning algorithms
- Significant improvement in computational power (**GPU**, distributed computing)
Working Ideas on Algorithms

Input

Feature engineering

End-to-end learning, *deep learning*

Result

classifier e.g., SVM

“dog”
Why has Deep Learning Been so Successful Lately?

- Large-scale annotated data sets (e.g., ImageNet, 14 million images in 22k categories)
- Deep learning algorithms
- Significant improvement in computational power (GPU, distributed computing)
Artificial Neural Networks

Training in progress...

Adaptable Weights

Weights updated with the Backpropagation Algorithm

Source: gfycat.com
ILSVRC’14 Winner: VGG-Net

- VGG-Net has 16/19 layers, 24M nodes, 14M parameters and, 15B connections
  - model size 550MB
  - memory: 24M * 4 bytes $\approx$ 96MB / image (only forward)

Simonyan et al. VGG-Net, ICLR’15
Why has Deep Learning Been so Successful Lately?

- Largescale annotated data sets (e.g., ImageNet, 14 million images in 22k categories)
- Deep learning algorithms
- **Significant improvement in computational power** (GPU, distributed computing)
Computational Power

Rapid development of hardware acceleration and massive amounts of computational power

- Applying GPUs/TPUs in neural network computation
- Training time of a very deep model:
  10 years ago: several months → Today: ?
- Cloud computing, distributed system

Yet Another Accelerated SGD: ResNet-50 Training on ImageNet in 74.7 seconds

Masafumi Yamazaki, Akihiko Kasagi, Akihiro Tabuchi, Takumi Honda, Masahiro Miwa, Naoto Fukumoto, Tsuguchika Tabaru, Atsushi Ike, Kohta Nakashima

Fujitsu Laboratories Ltd.

{m.yamazaki, kasagi.akihiko, tabuchi.akihiro, honda.takumi, masahiro.miwa, fukumoto.naoto, tabaru, ike, nakashima.kouta}@fujitsu.com
Limitations of Deep Learning

• The main achievements are in supervised and reinforcement learning
  – Requiring more annotated data
  – Semi-supervised and weakly supervised methods do not perform well
• Computationally expensive
• Difficult to engineer with, architecture engineering
• Deep models have very limited interpretability
• Other issues such as adversarial attack, ethical issue, inability to distinguish causation from correlation, not well being integrated with prior knowledge, and other potential risks
Research Questions

• **Q1**: How can we alleviate the reliance on substantial data annotations of DL?
  – Through synthetic data?
  – Through unsupervised or semi-supervised learning method?

• **Q2**: How can we perform multiple computer vision tasks with a uniform end-to-end neural network?

• **Q3**: How can we apply DL models on low power devices as e.g., smart phones, embedded devices?

• **Q4**: Can DL models gain multimodal and cross-modal representation learning tasks?

• **Q5**: Can we effectively apply multimedia analysis and DL algorithms in real-world applications?

Publications

• During my Ph.D. study (2010-2013): 13 papers
  – Ph.D. thesis: *automatic video indexing and retrieval using video OCR technology* (summa cum laude)

• After Ph.D. (2014-preset): > 45 papers

Q1.1: “SceneTextReg”

Q1.2, Q2: “SEE”

Q3: “BMXNet”

Q4: “Neural Captioner”

Q5: “Automatic Online Lecture Highlighting”
  “Medical Image Segmentation”
Selected Publications

- SceneTextReg: A real-time video ocr system, ACM Multimedia 2016
- SEE: Towards semi-supervised end-to-end text recognition, AAAI 2018
- BMXNet
  - Bmxnet: An open-source binary neural network implementation based on mxnet, ACM Multimedia 2017
  - Back to simplicity: How to train accurate BNNs from scratch? ICCV 2019 (under review)
- Neural Captioner: Image captioning with deep bidirectional LSTMs and multi-task learning, ACM Trans. Multimedia Computing 2018
- RE-DNN: A deep semantic framework for multimodal representation learning, Multimedia Tools and Applications 2016
- Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation, Multimedia Tools and Applications 2019
- Automatic online lecture highlighting based on multimedia analysis, IEEE Trans. Learning Technology 2018
Selected Publications

- **SceneTextReg**: A real-time video ocr system, ACM Multimedia 2016
- **SEE**: Towards semi-supervised end-to-end text recognition, AAAI 2018
- **BMXNet**
  - Bmxnet: An open-source binary neural network implementation based on mxnet, ACM Multimedia 2017
  - Back to simplicity: How to train accurate BNNs from scratch? ICCV 2019 (under review)
- **Neural Captioner**: Image captioning with deep bidirectional LSTMs, ACM Trans. Multimedia Computing, 2018
- **RE-DNN**: A deep semantic framework for multimodal representation learning, Multimedia Tools and Applications 2016
- **Automatic online lecture highlighting based on multimedia analysis**, IEEE Trans. Learning Technology 2018
Print OCR vs. OCR in Multimedia

SET encoding
Set character encoding of words and morph
UTF-8, ISO8859-1 – ISO8859-10, ISO885
CP1251, ISCII-DEVANAGARI.

FLAG value
Set flag type. Default type is the extended AS
encoded Unicode character flags. The ‘long’
type, the ‘num’ sorts the decimal number flag
in flag fields are separated by comma. BUG:

COMPLEX PREFIXES
SceneTextReg

SceneTextReg: real-time scene text recognition, Yang, Wang, Bartz and Meinel, ACM MM’16

Matas et al. Maximally Stable Extremal Regions, BMVC’02
SceneTextReg - Data Generator

Features

• Various fonts (>1500)
• Different colors, sizes, shadows, borders with varying displacements to the rendered texts
• Transformations: distortion, rotation
• Random blur, reflection
• Background blending (nature scene images)

ICDAR data set (right column)
SceneTextReg - Evaluation

62-way char classification (on ICDAR'03 data set):

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our result</td>
<td>0.872</td>
</tr>
<tr>
<td>Jaderberg et al. (ECCV’14)</td>
<td>0.868</td>
</tr>
<tr>
<td>Alsharif et al. (ICLR’14)</td>
<td>0.86</td>
</tr>
<tr>
<td>Wang et al. (ICPR’12)</td>
<td>0.839</td>
</tr>
<tr>
<td>A. Coates et al. (ICDAR’11)</td>
<td>0.817</td>
</tr>
</tbody>
</table>

ICDAR’13/15 data set (IAPR International Conference on Document Analysis and Recognition) on focused scene word recognition:

<table>
<thead>
<tr>
<th>Method</th>
<th>WRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google’s PhotoOCR</td>
<td>0.8283</td>
</tr>
<tr>
<td>SceneTextReg</td>
<td>0.8237</td>
</tr>
<tr>
<td>PicRead</td>
<td>0.5799</td>
</tr>
<tr>
<td>NESP</td>
<td>0.642</td>
</tr>
<tr>
<td>PLT</td>
<td>0.6237</td>
</tr>
<tr>
<td>MAPS</td>
<td>0.6274</td>
</tr>
<tr>
<td>PIONEER</td>
<td>0.537</td>
</tr>
<tr>
<td>ABBY OCR SDK10</td>
<td>0.453</td>
</tr>
</tbody>
</table>

Only synthetic data used for training!

WRA: Word Recognition Accuracy
*(Case sensitive with punctuation, special chars)*

Bissacco et al. PhotoOCR, ICCV’13
SceneTextReg - Demo
Scene Text Recognition with NN

- Two stage system as e.g., *SceneTextReg*

- End-to-end system as e.g., *Faster RCNN*
SEE: Towards Semi-Supervised End-to-End Scene Text Recognition, Bartz, Yang, Meinel, AAAI 2018
SEE - Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxout CNN, (ICLR'14)</td>
<td>0.96</td>
</tr>
<tr>
<td>ST-CNN, (NIPS'15)</td>
<td>0.963</td>
</tr>
<tr>
<td>SEE</td>
<td>0.952</td>
</tr>
</tbody>
</table>

SVHN house number data set

<table>
<thead>
<tr>
<th>Method</th>
<th>IC13/15</th>
<th>SVT</th>
<th>IIIT5K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google's PhotoOCR, (ICCV'13)</td>
<td>0.876</td>
<td>0.78</td>
<td>-</td>
</tr>
<tr>
<td>CharNet, (ECCV'14)</td>
<td>0.818</td>
<td>0.717</td>
<td>-</td>
</tr>
<tr>
<td>CRNN, (TPAMI'16)</td>
<td>0.867</td>
<td>0.808</td>
<td>0.782</td>
</tr>
<tr>
<td>RARE, (CVPR'16)</td>
<td>0.875</td>
<td>0.819</td>
<td>0.819</td>
</tr>
<tr>
<td>SEE</td>
<td><strong>0.903</strong></td>
<td>0.798</td>
<td><strong>0.86</strong></td>
</tr>
</tbody>
</table>

ICDAR’13/15, SVT, IIIT5K data set
## SEE - Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith et al. (Google) (ECCV’16)</td>
<td>0.725</td>
</tr>
<tr>
<td>Wojna et al. (Google) (ICDAR’17)</td>
<td>0.842</td>
</tr>
<tr>
<td>SEE</td>
<td>0.78</td>
</tr>
</tbody>
</table>

French street name signs data set
SEE - Demo
Multimodal Retrieval

- Image captioning
- Video classification
- Human action recognition in surveillance video
Neural Captioner

*Image Captioning with Deep Bidirectional LSTMs, Wang, Yang, Bartz and Meinel, ACM MM’16*

- Visual representation → CNN model
  - Transfer learning from ImageNet models
- Visual to sentence (language) embedding
  - **Bi-directional LSTM** (Long Short-Term Memory)
- Data augmentation: random cropping, mirroring, shifting
Neural Captioner

The proposed architectures

• baseline model (a)
• bidirectional LSTM (b)
• bidirectional Stacked LSTM (c)
• bidirectional LSTM with fully connected (FC) transition layer (d)
Neural Captioner

Contributions

- Cover more semantics by Bi-LSTM
- Great portion of generated sentences not appear in training set
- Achieved state-of-the-art on Flickr8K, Flickr30K, MSCOCO and Pascal1K image captioning data sets
Neural Captioner - Demo
Deep Learning on Low Power Devices

A state-of-the-art ResNet-152 (152 layers) surpasses human performance on the image classification task.

Number of operations:
- AlexNet (240MB), 720 MFLOPs,
- VGG19 (550MB), 19.6 BFLOPs
- ResNet-152 (240MB), 11.3 BFLOPs

Inference time on CPU:
- AlexNet: 3 fps,
- VGG19: 0.25 fps
- ResNet-152: 0.63 fps
Deep Learning on Mobile Devices

Autonomous driving

Assistance apps

Low power devices
**Binary Neural Networks**

**Benefits**

- 32x smaller model size
  - e.g., FPGAs with <10MB on-ship memory
- 32x less memory access → much less energy consumption
- Bitwise operator e.g., XNOR, bitcount instead of arithmetic operations in NN
  - It allows for a speedup factor of up to 32 by combining multiple operations in one CPU cycle
- On devices, offline prediction → better privacy protection
BMXNet
An open-source binary neural network implementation based on mxnet, Yang, Fritzsche, Bartz and Meinel, ACM MM’17

- Flexible design and fully compatible with standard neural network components
- Source code: https://github.com/hpi-xnor
- E.g., ResNet-18 for image classification on Cifar-10 data set
  - 45MB (full precision) → 1.5MB (binary)
Contributions:

- Challenging conventional wisdom: Highly accurate BNNs can be trained by using standard training strategy.
- We suggest general design principles for BNNs.
- Our BinaryDenseNet significantly surpasses all existing BNNs for image classification without tricks.
- We provide codes to facilitate follow-up studies.
BMXNet - Evaluation

<table>
<thead>
<tr>
<th>Model size</th>
<th>Method</th>
<th>Top-1/Top-5 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>~4.0MB</td>
<td>XNOR-ResNet18 (ECCV’16)</td>
<td>51.2%/73.2%</td>
</tr>
<tr>
<td></td>
<td>TBN-ResNet18 (ECCV’18)</td>
<td>55.6%/74.2%</td>
</tr>
<tr>
<td></td>
<td>Bi-Real-ResNet18 (ECCV’18)</td>
<td>56.4%/79.5%</td>
</tr>
<tr>
<td></td>
<td>BinaryResNetE18 (ours)</td>
<td>58.1%/80.6%</td>
</tr>
<tr>
<td></td>
<td>BinaryDenseNet28 (ours)</td>
<td>60.7%/82.4%</td>
</tr>
<tr>
<td>~5.1MB</td>
<td>TBN-ResNet34 (ECCV’18)</td>
<td>58.2%/81.0%</td>
</tr>
<tr>
<td></td>
<td>Bi-Real-ResNet34 (ECCV’18)</td>
<td>62.2%/83.9%</td>
</tr>
<tr>
<td></td>
<td>BinaryDenseNet37 (ours)</td>
<td>62.5%/83.9%</td>
</tr>
<tr>
<td></td>
<td>BinaryDenseNet37-dilated (ours)</td>
<td>63.7%/84.7%</td>
</tr>
<tr>
<td>7.4MB</td>
<td>BinaryDenseNet45 (ours)</td>
<td>63.7%/84.8%</td>
</tr>
<tr>
<td>46.8MB</td>
<td>Full-precision ResNet18</td>
<td>69.3%/89.2%</td>
</tr>
<tr>
<td>249MB</td>
<td>Full-precision AlexNet</td>
<td>56.6%/80.2%</td>
</tr>
</tbody>
</table>

The trade-off of top-1 validation accuracy on ImageNet and number of operations. All the binary/quantized models are based on ResNet18 except BinaryDenseNet.
BMXNet - Demo
Thank you for your Attention!

0 to 3

“Medical Image Segmentation”

“Automatic Online Lecture Highlighting”

“SEE”

“Neural Captioner”

“BMXNet”

“SceneTextReg”

Beyond!