Modern Technology is Awesome!

- Pack 7 billion transistors on a single chip;
- Routinely fly commercial passenger aircraft at 70% the speed of sound;
- Construct buildings ½ mile tall;
- 5 billion videos are watched on YouTube every day;
- Trade 2 billion shares per day;
- Lab-grown body parts are being used for life-saving transplants.
Evolution of Change

- 30 years ago computers were for data entry and simple processing.
- 20 years ago cell phones were for voice and very limited texting.
- 10 years ago smart phones were primitive, buggy, and expensive.
- 5 years ago deep learning was slow and ineffective.

Ph.D. Students Need to Change as Well!

- Year 1: So hungry for knowledge, wanted to work on everything. 1.5
- Year 2: Lots of applications oriented research-Wanted to build things to change the world! 2.5
- Year 3: Going after top-tier venues. Reviewers are brutal! Adding more algorithms and math. 3
- Year 4: Reviewers don’t care about applications, they want math/science contributions. 4
- Year 5: Build on foundation. Focus on publishing. 9
• Unless you are hungry for knowledge, you should not be getting a Ph.D.
• Listen to your advisor on both research topics and how to write papers.
• Find a topic you have passion for - stay focused!

Your Ph.D. is a 24/7 job…
• If you are not feeling challenged/stressed, you are not working hard enough.
• Don’t work in a vacuum, actively seek out advice from your peers and experts in the field.
• Be proud of everything you do.

A Little Background…
Impact of Machine Learning

- Machine learning is giving computers the ability to analyze, generalize, think/reason/behave like humans.
- Machine learning is transforming medical research, financial markets, international security, and generally making humans more efficient and improving quality of life.
- Inspired by the mammalian brain, deep learning is machine learning on steroids- bigger, faster, better!
Machine Learning is a Hot Topic!

- Machine learning, cs229 is most popular course at Stanford.
- Deep learning class, cs231 went from 150 in 2015 to 350 in 2016 to 750 in 2017!

AI Trends

- Supervised learning has generated billions of marketable products- targeted advertising, click through rates on web pages, and driver assistance are just a few examples.
- Despite all the AI hype today, humans are still much more capable than machines at general tasks.

..however, for targeted tasks, things are different!

- Rule of thumb: Anything that a human can do with ≤ one second of thought can be probably now or soon be automated with AI.
Deep Learning
Hottest topic in pattern recognition

Vision
“Calista Flockhart”

Speech
“Greetings ladies and gentlemen”

Network security
Medical diagnosis
Financial markets

Vision Tasks

Classification
Classification + Localization
Object Detection
Instance Segmentation

CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Speech Recognition

- Conversion of speech to spectrograms, then from spectrograms to words.


OCR/ICR Tasks
Language Translation

• In 2014, Sutskever, Vinyals, and Le (from Google) showed that a simple encoder-decoder framework was just as good as sophisticated Statistical Machine Translation (SMT) systems, and almost as good as SMT systems paired with nnets.

Sutskever et al., NIPS ’14

Image/Video Captioning

[Fang et al. CVPR15] [Vinyals et al. CVPR15]

[A group of people shopping at an outdoor market.]
[A woman holding a camera in a crowd.]

Ptucha ’17
Two Most Important Deep Learning Fields

• Convolutional Neural Networks (CNN)
  – Examine high dimensional input, learn features and classifier simultaneously

• Recurrent Neural Networks (RNN)
  – Learn temporal signals, remember both short and long sequences
Many Flavors of CNNs…

LeNet-5, LeCun 1989

AlexNet, Krizhevsky 2012

VGGNet, Simonyan 2014

GoogLeNet (Inception), Szegedy 2014

ResNet, He 2015

DenseNet, Huang 2017

Image Convolution

By padding (filterWidth-1)/2, output image size matches input image size

input

output

3x3 filter sliding over input image

Vert pad

Horiz pad

https://github.com/vdumoulin/conv_arithmetic
Max Pooling- Reducing the Size of an Image

<table>
<thead>
<tr>
<th>Single depth slice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

max pool with 2x2 filters and stride 2

<table>
<thead>
<tr>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

cs321n, Karpathy, Li

Convolution Neural Network (CNN) Building Block

Pooling

Convolution

Image

Deng ICML '14
Putting it All Together

Whole System

Convolution → Pooling

Input Image

1st stage 2nd stage 3rd stage

Fully Conn. Layers → Class Labels

Learning Filters

32 Learned filters, each 5×5

Input image 28×28

32 Filtered images (activation maps), each is 28×28

Use zero padding

Ptucha '17
Learning Filters

32 Learned filters, each $5 \times 5 \times 3$

Input image
$28 \times 28 \times 3$

32 Filtered images, each is $28 \times 28 \times 1$

Use zero padding

Ptucha '17 37

CNN Architecture

(Not so) Toy Example

Input RGB image:
$64 \times 64 \times 3$ pixels

Max pooling $\times 2$

16 filters, each filter is $5 \times 5 \times 32$.
2 pixel pad.

Max pooling $\times 2$

32 filters, each filter is $5 \times 5 \times 16$.
2 pixel pad.

Max pooling $\times 2$

64 filters, each filter is $5 \times 5 \times 32$.
2 pixel pad.

$1 \times 1 \times 64$.
filter, 0 pixel pad.

4$\times$4 converted to 16 element vector

Output:
prediction of 1 of 10 categories

Fully connected to 10 classes

Ptucha '17 38
CNN Visualization

...amazingly, these filters form an abstract hierarchy...

Zeiler, Fergus, 2014

Ptucha '17 40

CNN Visualization

Layer 1

Layer 2

Layer 4

Layer 5

Zeiler, Fergus, 2014

Ptucha '17 41
CNN Results

- Handwritten characters
  - MNIST: 0.17% error, Ciresan et al. ’11
  - Arabic & Chinese: Ciresan ‘12

- CIFAR-10 (60K images of 10 classes)
  - 9.3% error, Wan et al. ‘13

- Traffic Sign Recognition
  - 0.56% error vs 1.16% for humans, Ciresan ‘11

ImageNet

- Amazon Turk did bulk of labeling
- 14M labeled images
- 20K classes

- 1.2M images, 1000 categories
- Image classification, object localization, video detection
ImageNet: Examples of Hammer

Deep Learning - Surpassing The Visual Cortex’s Object Detection and Recognition Capability

Top-5 error on ImageNet

- Traditional Computer Vision and Machine Learning
- Deep Convolution Neural Networks (CNNs)

Introduction of deep learning

Trained Human (genius intellect)

Similar effect demonstrated on voice and pattern recognition

Year

Error


Ptucha '17
Two Most Important Deep Learning Fields

- Convolutional Neural Networks (CNN)
  - Examine high dimensional input, learn features and classifier simultaneously

- Recurrent Neural Networks (RNN)
  - Learn temporal signals, remember both short and long sequences

Artificial Neuron

Note, $x_0$ is the bias unit, $x_0=1$

$h_\theta(x) = g(x_0\theta_0 + x_1\theta_1 + \ldots + x_n\theta_n) = g \left( \sum_{i=0}^{n} x_i\theta_i \right)$

$h_\theta(x) = g(\theta^T x)$
Artificial Neural Networks

• Artificial Neural Network (ANN) – A network of interconnected nodes that “mimic” the properties of a biological network of neurons.

Adding Recurrence

\[ z = \sigma (x_0 \theta_0 + x_1 \theta_1 + ... + x_n \theta_n) = \sigma (\theta^T x) \]
Recurrent Networks

\[ i_{h0} = (W_{xh}x_t + W_{hh}h_{t-1}) \]
\[ h_t = f(i_{h0}) \]
\[ i_{y0} = W_{hy}h_t \]
\[ y_t = f(i_{y0}) \]

Where:
- \( f \) is some activation function
- \( x_t, h_t, h_{t-1} \) and \( y_t \) are current input, hidden, previous hidden and current output values
- \( W_{xh}, W_{hh} \) and \( W_{hy} \) are the weight matrices for input, hidden and output stages respectively
- \( i_{h0} \) and \( i_{y0} \) are the inputs to activation function in hidden and output layers
Recurrent Networks
Both figures represent the same architecture

---

Recurrent Networks
Recurrent Neural Network “neuron”

P(next event | previous events)

- Unfortunately, these vanilla RNNs don’t always work.
- Can’t store info over long periods of time.
- Suffer from vanishing and/or exploding gradients.
Recurrent Networks

- LSTM's allow read/write/reset functions to neurons.
- Remember past to predict the future- (over long time periods).
- Can have many hidden neurons per layer and many layers.

Recurrent Applications

Sutskever et al., 2014
Learning Shakespeare

• LSTMs can learn structure and style in the data
• Karpathy downloaded all the works of Shakespeare and concatenated them into a single (4.4MB) file.
• Train a 3-layer LSTM with 512 hidden nodes on each layer.
• After we train the network for a few hours Karpathy obtained samples such as:

```
PTUCHA '17 59
http://karpathy.github.io/2015/05/21/rnn-effectiveness/
```

```plaintext
PANDARUS:
Alas, I think he shall be come approached and the day
When little strain would be attain'd into being never Fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.
```
Learning LaTeX

• The results above suggest that the model is actually quite good at learning complex syntactic structures.
• Karpathy and Johnston downloaded the raw Latex source file (a 16MB file) of a book on algebraic stacks/geometry and trained a multilayer LSTM.
• Amazingly, the resulting sampled LaTex almost compiled.
• They had to step in and fix a few issues manually but then they get plausible looking math:
Two Most Important Deep Learning Fields

• Convolutional Neural Networks (CNN)
  – Examine high dimensional input, learn features and classifier simultaneously

• Recurrent Neural Networks (RNN)
  – Learn temporal signals, remember both short and long sequences

• Combining Visual CNNs with Language RNNs for Image Captioning
Image Captioning

CNN helps represent an image as a numeric value. (image2vec)

RNN takes in a latent representation of an image, and generates a sequence.

- We may have 50K words. Instead of one-hot encoding, we learn an embedding for each word.
- Glove embedding (300 long vector/word) is very popular.
- Alternately, can learn embedding - learn a matrix which goes from (50K) one-hot to 300, ie: \( W_{ix} \in \mathbb{R}^{50K \times 300} \)
- Embedding and unembedding can be learned or inverses of one another.
Note: Word is sampled from distribution of word probabilities

\[ y_t = f(W_{hy}h_t) \]

\[ h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + W_{vh}v) \]

\[ v \] could be FC6, FC7, conv5, conv4, …; or a combination of above

Training samples are:
\[ <word1>, <word2>, \ldots <wordn>, <EOS> \]
While word embedding is 300, \( x \in \mathbb{R}^{300} \), the hidden embedding can be anything, such as 512.
Beam Size Example

Beam size = 3

Increased probability

Only keep top three sentences at each time step

PTUCHA '17 70

Beam Size Example

Beam size = 3

PTUCHA '17 71
Beam Size Example

T = 0  T = 1  T = 2  T = 3  T = 4  T = 5

Beam size = 3

Only keep growing top three sentences at each time step

Ptucha '17 72

Beam Size Example

T = 0  T = 1  T = 2  T = 3  T = 4  T = 5

Beam size = 3

Only keep growing top three sentences at each time step

Ptucha '17 73
Beam Size Example

Beam size = 3

Only keep growing top three sentences at each time step

Beam Size Example

Beam size = 3

Only keep growing top three sentences at each time step
Data for Captioning

- Flickr8K
  - 8,000 images, from Flickr website, each with five captions
  - http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html
- Flickr30K
  - 31,783 images, from Flickr website, each with five captions
  - http://shannon.cs.illinois.edu/DenotationGraph/
- MSCOCO
  - 80,000 training images, each with five captions
  - http://mscoco.org/
Captioning Datasets

Amazon mechanical turkers do all labeling
https://www.mturk.com/mturk/welcome

MSCOCO Dataset

A pile of wooden boxes filled with fruits and vegetables.
An assortment of fruit in buckets for sale in a shop.
An outdoor fruit stand with various types of fruits for sale.
A display of crates of fruit on a city street.
There are many crates with fruit and vegetables.

A cat stands on a counter while a dog stands on the floor.
A cat on the kitchen counter is looking down at a dog.
A cat is looking at a dog rummaging in the garbage.
A cat on the counter and a dog on the ground in the kitchen.
A cat stalking a dog on the kitchen floor.
Video Data for Captioning

MSVD: Microsoft Video Description Dataset
MSR-VTT: Microsoft Research -Video to Text
M-VAD: Movie description dataset M-VAD

<table>
<thead>
<tr>
<th></th>
<th>MSVD</th>
<th>MSR-VTT</th>
<th>M-VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sentences</td>
<td>80,827</td>
<td>200,000</td>
<td>54,997</td>
</tr>
<tr>
<td>#sent. per video</td>
<td>~42</td>
<td>20</td>
<td>~1-2</td>
</tr>
<tr>
<td>vocab. size</td>
<td>9,729</td>
<td>24,282</td>
<td>16,307</td>
</tr>
<tr>
<td>avg. length</td>
<td>10.2s</td>
<td>14.8s</td>
<td>5.8s</td>
</tr>
<tr>
<td>#train video</td>
<td>1,200</td>
<td>6,513</td>
<td>36,921</td>
</tr>
<tr>
<td>#val. video</td>
<td>100</td>
<td>497</td>
<td>4,651</td>
</tr>
<tr>
<td>#test video</td>
<td>670</td>
<td>2,990</td>
<td>4,951</td>
</tr>
</tbody>
</table>
The Novel Object Captioner (NOC)

- Visual network (left), language network (right), and caption network (center) are all trained simultaneously with different objectives, but with shared parameters.
The Machine Intelligence Lab

- Advanced deep learning research-developing new algorithms and deep architectures.
- Research on tomorrow’s smart devices: Machine learning, robotics, HCI, computer vision, NLP, …
- Teaching a robot how to recognize objects, interact with people, or navigate complex environments.
- All students experts in deep learning.
**Machine Intelligence Lab**

- Mostly masters students in computer engineering:
  - Includes 2+2+1/2 PhD students
  - Includes students from computer science, electrical engineering, imaging science, and computer engineering
- 34 publications since 2015
- 2 patent applications
- Deep learning tutorials, consultation, online classes, reading groups.

---

**Summarization of Long Videos**: Tradition computer vision is used to detect interesting regions of video, an encoder-decoder uses CNN representation of frames and kicks out textual descriptions. Traditional text summarizes combine all descriptions into paragraphs.

**Semantic Text Summarization of Long Videos**

**Introduction**
- Our work proposes methods to generate both visual and textual summaries of long videos.
- Videos of several hours long are passed into a superpixel segmentation algorithm.
- Each superpixel segmentation is evaluated for interestingness (boundary motion, superpixel motion, aspect ratio, motion, sharpness, contrast, and facial stereo).
- The superpixel module receives temporal segments, and generates captions.
- Captions are then passed into the summarization task, which generates a single sentence per paragraph per video.

**Super-Segment Selection**

- Interestingness is determined by a non-linear combination of scores measuring boundary motion, contrast, sharpness, distortion, and facial stereo.

**Dataset**

- Vidyafett dataset with 10 daily-life episodes. Binary-world-oppositional TV episodes with distemper between 2.3 and 1.0 and episode for every 5-10 seconds. We then take the caption on 5 videos and text on remaining 3 videos.

---

**Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Video</th>
<th>Train Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vidyafett</td>
<td>114.5</td>
<td>128.2</td>
</tr>
<tr>
<td>Binary World</td>
<td>166.3</td>
<td>180.4</td>
</tr>
<tr>
<td>Dynamic</td>
<td>180.4</td>
<td>194.5</td>
</tr>
</tbody>
</table>

**Sample Positive Generated Summary**

- Summary length (words) to be in range of 100 words.
- Captions should be informative and coherent.
- Captions should be descriptive and detailed.
- Captions should be concise and focused.
- Captions should be visually appealing and engaging.
- Captions should be meaningful and relevant.
- Captions should be interesting and engaging.
-Captions should be well-structured and organized.
**Advanced Video Understanding:** This paper combines several concepts: Gaussian attention such that input to attention can be varying length; attention steering features for richer steering; global video features; and hierarchical LSTM.

**Temporally Steered Gaussian Attention for Video Understanding**

**Ptucha '17**

---

**Most of the World’s Problems Not Homogeneous**

**Graph-CNN, Petroski Such '17**

**Ptucha ‘17**
More Info: Upcoming Talks

• AI Seminar Series: “Connecting Vision & Language”
  – Mon, Sept 11, 12:20-1:15pm, Bamboo Room (2nd floor off of SAU)
  – Overlap with this talk, but goes into details of latest MIL research

• Brick City Homecoming- “What Is And How Will Machine Learning and Artificial Intelligence Change Our Lives”
  – Sat, Oct 14, 10:30-11:30am, 12:30-1:30pm
  – Intro to machine and deep learning, discuss impact of AI on our lives

For More Information:  [http://www.rit.edu/mil](http://www.rit.edu/mil)