Connecting Vision and Language

Raymond Ptucha,
Rochester Institute of Technology

~ GCCIS PhD Colloquium ~
~ Rochester Joint Chapter of the IEEE Computer and Computational Intelligence Societies ~

Sept 08, 2017
GOL-1610

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.

Burj Khalifa in Dubai- 2,717 ft tall

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.

Ptucha '17 5

Ph.D. Students Need to Change as Well!

Ph.D.
Doctor of Philosophy
Computing and Information Sciences

- Year1: 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

Ptucha '17 6

Ph.D., Doctor of Philosophy, Golisano College of Computing and Information Sciences

- 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.

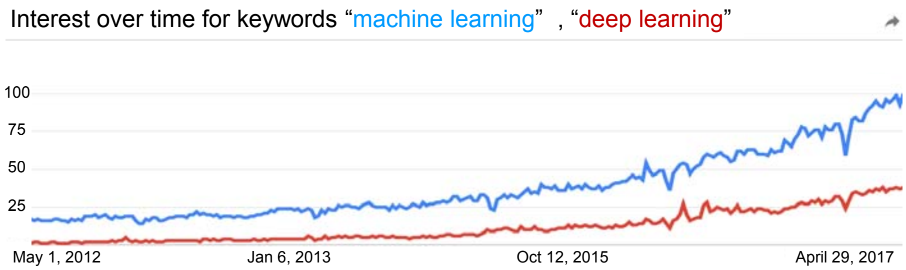
A Little Background…

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!



https://trends.google.com/trends/explore?q=machine%20learning

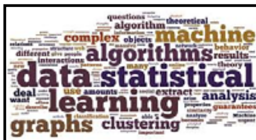
- Machine learning, cs229 is most popular course at Stanford.
- Deep learning class, cs231 went from 150 in 2015 to 350 in 2016 to 750 in 2017!

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.

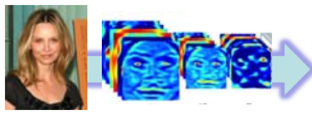
..however, for targeted tasks, things are different!

- Rule of thumb: Anything that a human can do with ≤ one second of thought can be probably now or soon be automated with AI.



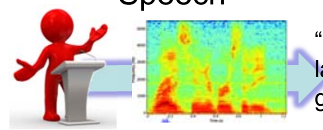
Deep Learning
Hottest topic in pattern recognition

Vision



"Calista Flockhart"

Speech



"Greetings ladies and gentlemen"

Network security

Medical diagnosis

Financial markets

Vision Tasks

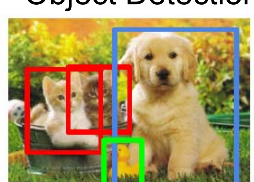
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Instance Segmentation

CAT, DOG, DUCK

Speech Recognition

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

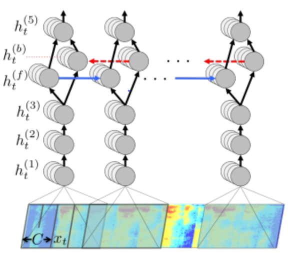
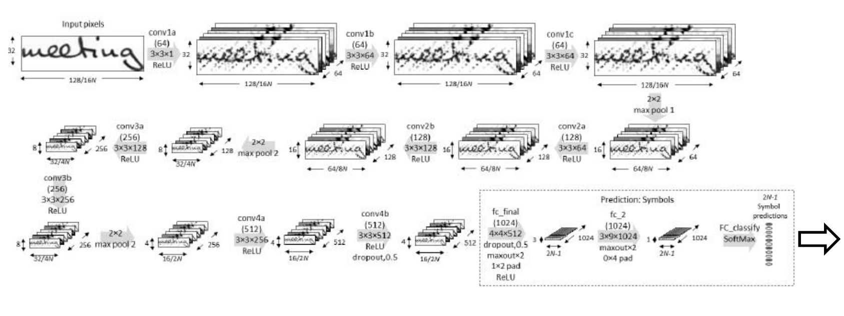


Figure 1: Structure of our RNN model and notation.

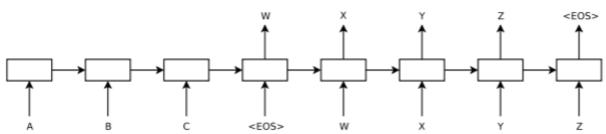
Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

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.

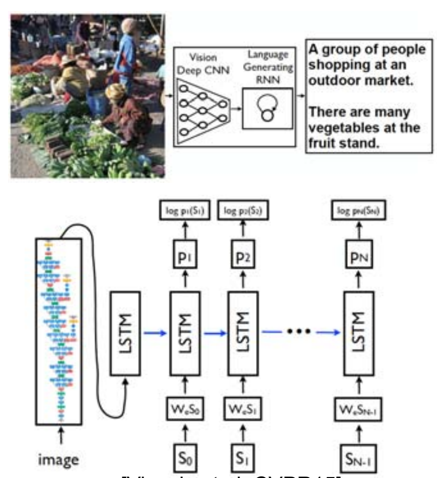


Sutskever et al., NIPS ‘14

Ptucha ‘17 21

Image/Video Captioning

[Fang et al. CVPR15]



[Vinyals et al. CVPR15]

Ptucha ‘17 22

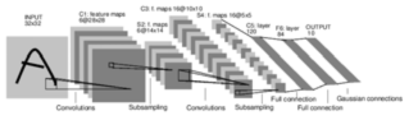
Two Most Important Deep Learning Fields

- 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

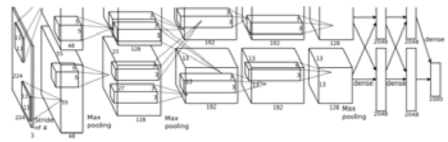
Two Most Important Deep Learning Fields

- 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

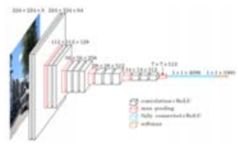
Many Flavors of CNNs…



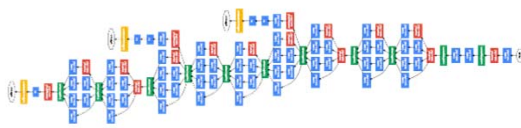
LeNet-5, LeCun 1989



AlexNet, Krizhevsky 2012



VGGNet, Simonyan 2014



GoogLeNet (Inception), Szegedy 2014

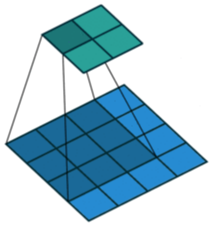
ResNet, He 2015

DenseNet, Huang 2017

Image Convolution

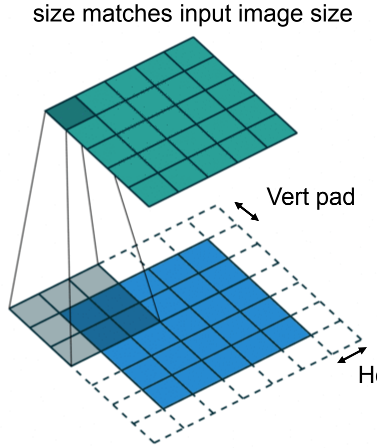
By padding (filterWidth-1)/2, output image size matches input image size

output



input

3×3 filter sliding over input image

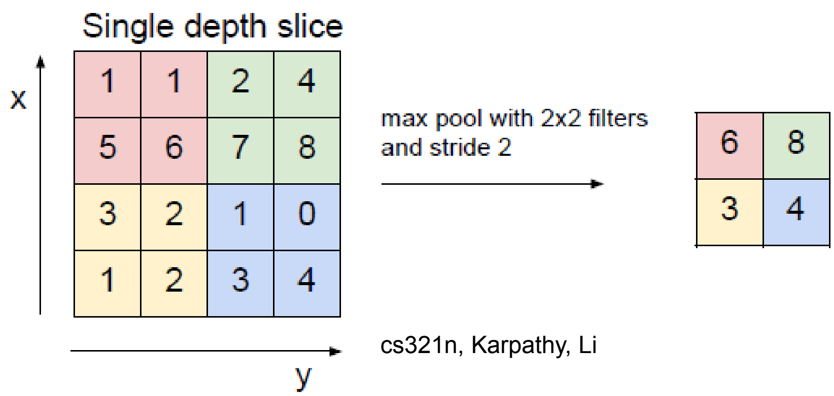


Vert pad

Horiz pad

https://github.com/vdumoulin/conv_arithmetic

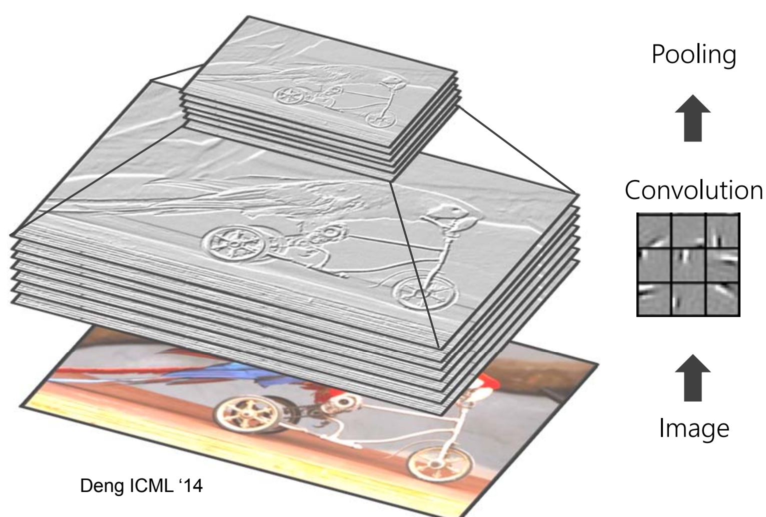
Max Pooling- Reducing the Size of an Image



cs321n, Karpathy, Li

Ptucha '17 31

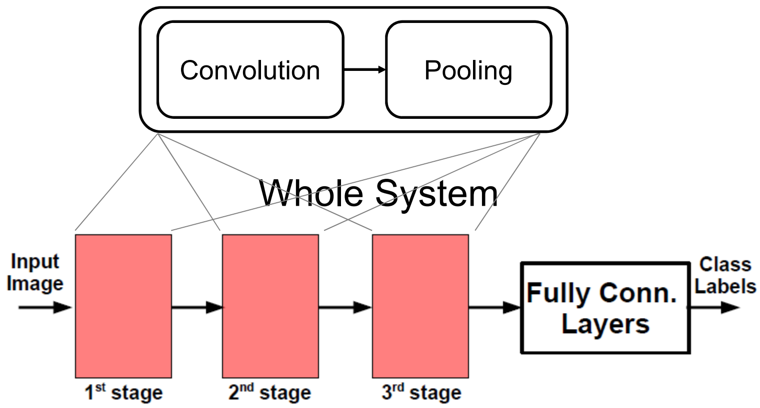
Convolution Neural Network (CNN) Building Block



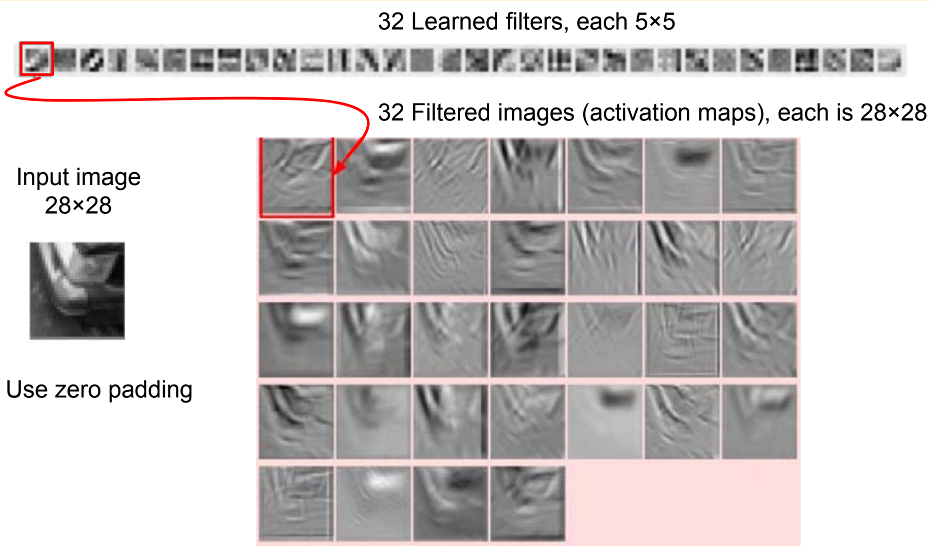
Deng ICML '14

Ptucha '17 32

Putting it All Together



Learning Filters

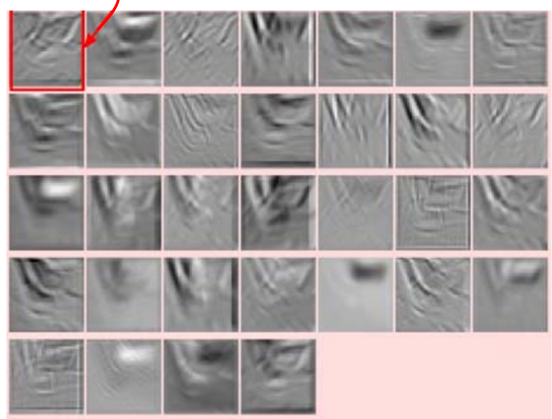


Learning Filters

32 Learned filters, each 5×5×3

32 Filtered images, each is 28×28×1

Input image
28×28×3



Use zero padding

CNN Architecture

(Not so) Toy Example

Output:
prediction of 1 of
10 categories

Input RGB image:
64×64×3 pixels

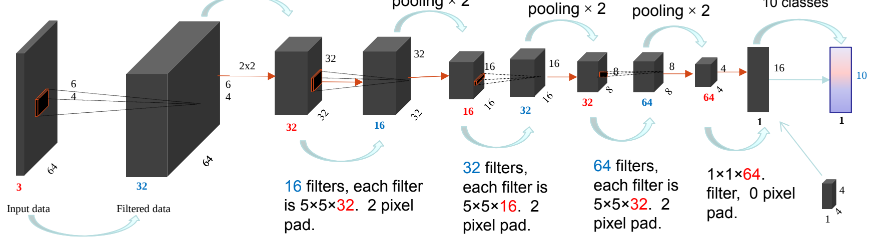
Max
pooling × 2

Max
pooling × 2

Max
pooling × 2

Max
pooling × 2

Fully
connected to
10 classes



32 filters, each filter is 5×5×3. 2 pixel pad added to top/bot/left/right so filtered image is same dimension as input image.

16 filters, each filter is 5×5×32. 2 pixel pad.

32 filters, each filter is 5×5×16. 2 pixel pad.

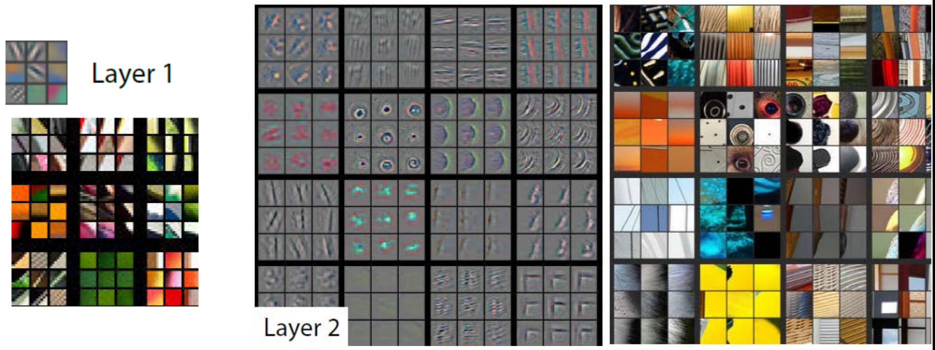
64 filters, each filter is 5×5×32. 2 pixel pad.

1×1×64. filter, 0 pixel pad.

4×4
converted
to 16
element
vector

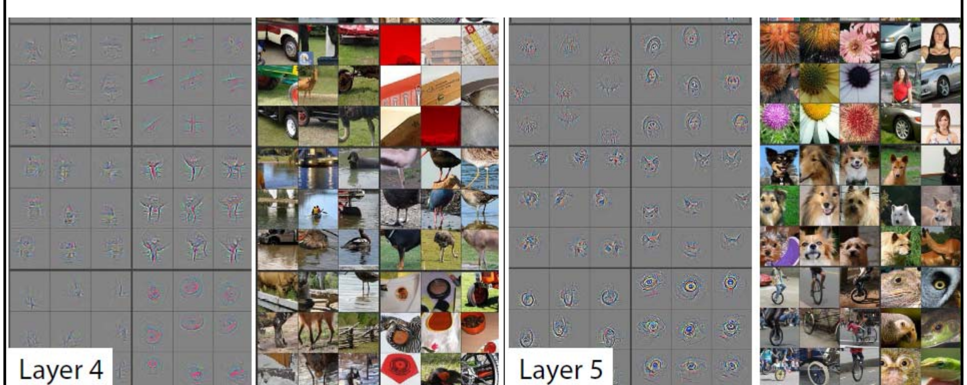
CNN Visualization

…amazingly, these filters form an abstract hierarchy…



Zeiler, Fergus, 2014

CNN Visualization



Zeiler, Fergus, 2014

CNN Results

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

CIFAR, Krizhevsky ‘09

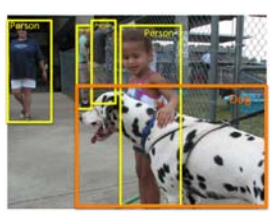
- CIFAR-10 (60K images of 10 classes)
  - 9.3% error, Wan et al. ‘13

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

Ciresan ‘11

ImageNet

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



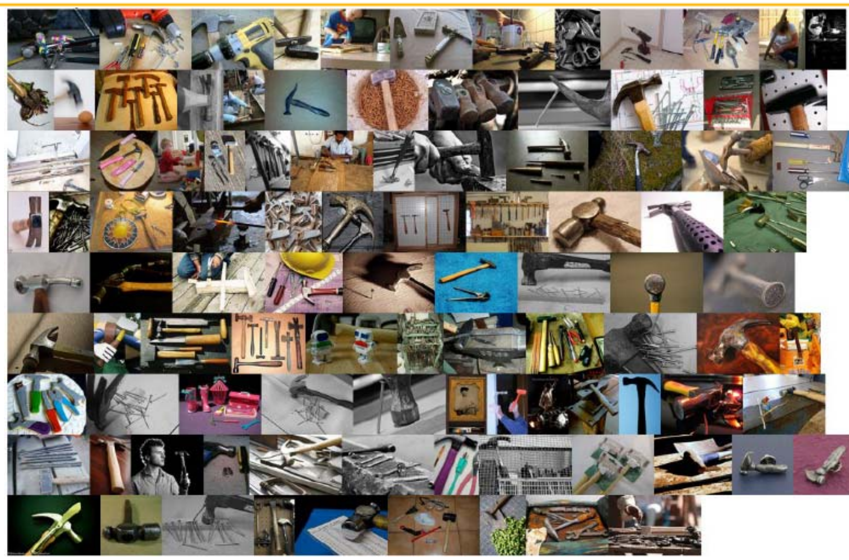
IMAGENET

Russakovsky et al., 2015

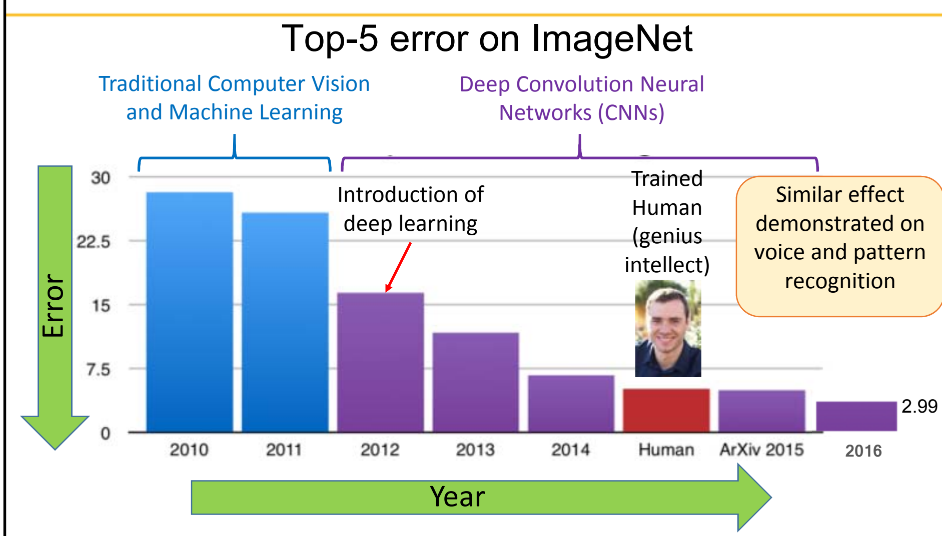
IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

- 1.2M images, 1000 categories
- Image classification, object localization, video detection

ImageNet: Examples of Hammer



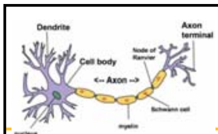
Deep Learning- Surpassing The Visual Cortex's Object Detection and Recognition Capability



Two Most Important Deep Learning Fields

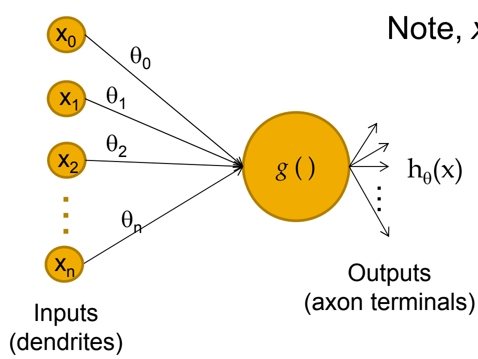
- Convolutional Neural Networks (CNN)
  - Examine high dimensional input, learn features and classifier simultaneously

- 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



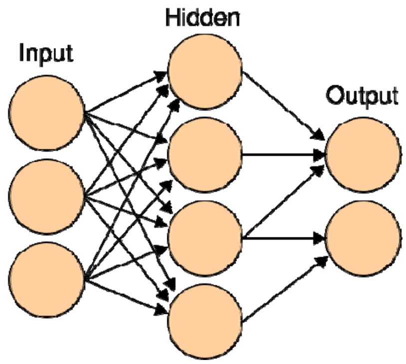
$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

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

$$h_\theta(x) = g(\theta^T x)$$

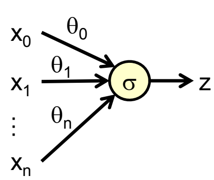
Artificial Neural Networks

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



Ptucha ‘17 49

Adding Recurrence

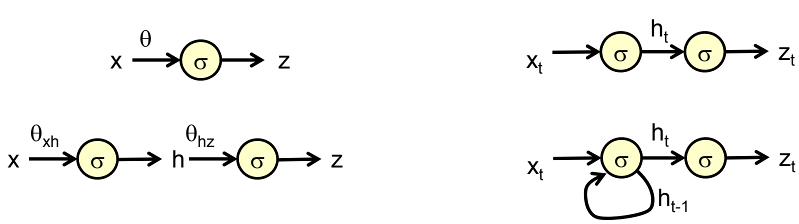


$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

Activation function

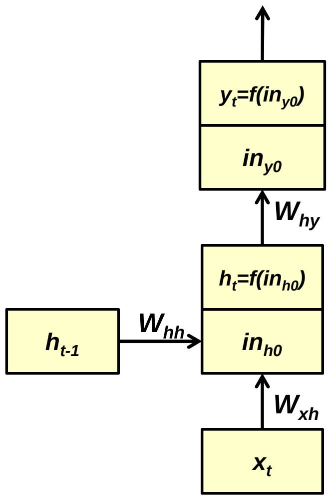
$$z = \sigma(x_0\theta_0 + x_1\theta_1 + \ldots + x_n\theta_n) = \sigma\left(\sum_{i=0}^{n} x_i\theta_i\right)$$

$$z = \sigma(\theta^T x)$$



Ptucha ‘17 50

Recurrent Networks

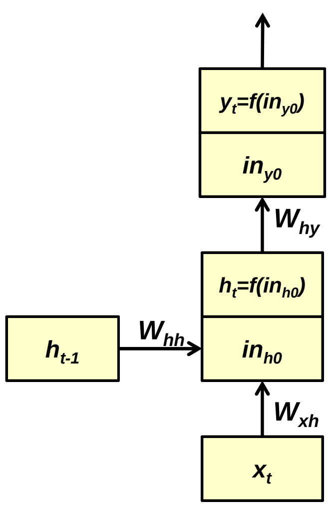


$$in_{h0} = (W_{xh}x_t \qquad\qquad)$$
$$h_t = f(in_{h0})$$

Where:

- f is some activation function
- x_t and h_t are current input and current output values
- W_{xh} is the weight matrix for input, hidden and output stages respectively
- in_{h0} is the input to activation function in hidden and output layers

Recurrent Networks



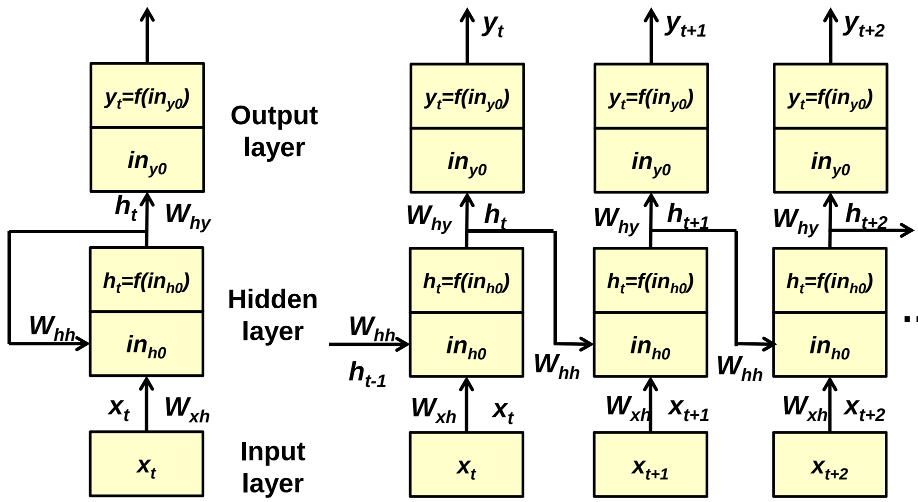
$$in_{h0} = (W_{xh}x_t + W_{hh}h_{t-1})$$
$$h_t = f(in_{h0})$$
$$in_{y0} = W_{hy}h_t$$
$$y_t = f\left(in_{y0}\right)$$

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
- in_{h0} and in_{y0} are the inputs to activation function in hidden and output layers

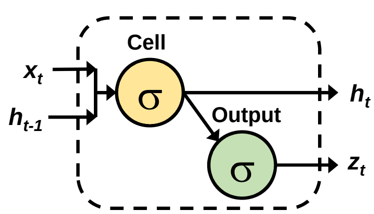
Recurrent Networks

Both figures represent the same architecture



Recurrent Networks

Recurrent Neural Network "neuron"

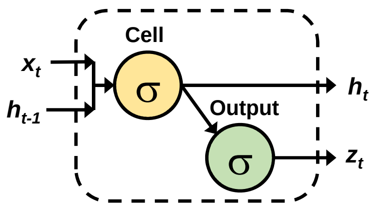


P(next event | previous events)

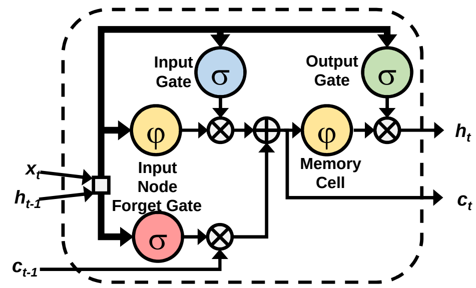
- Unfortunately, these vanilla RNNs don't always work.
- Can't store info over long periods of time.
- Suffer from vanishing and/or exploding gradients.

Recurrent Networks

Recurrent Neural Network "neuron"



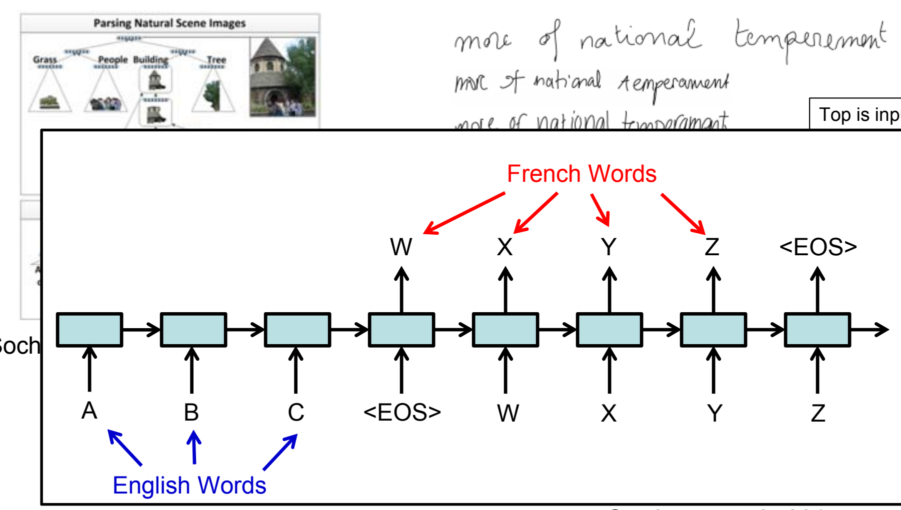
Long Short Term Memory "neuron"



Donahue et al., 2015

- 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



Soch

Top is input,

Sutskever et al., 2014

Learning Shakespeare

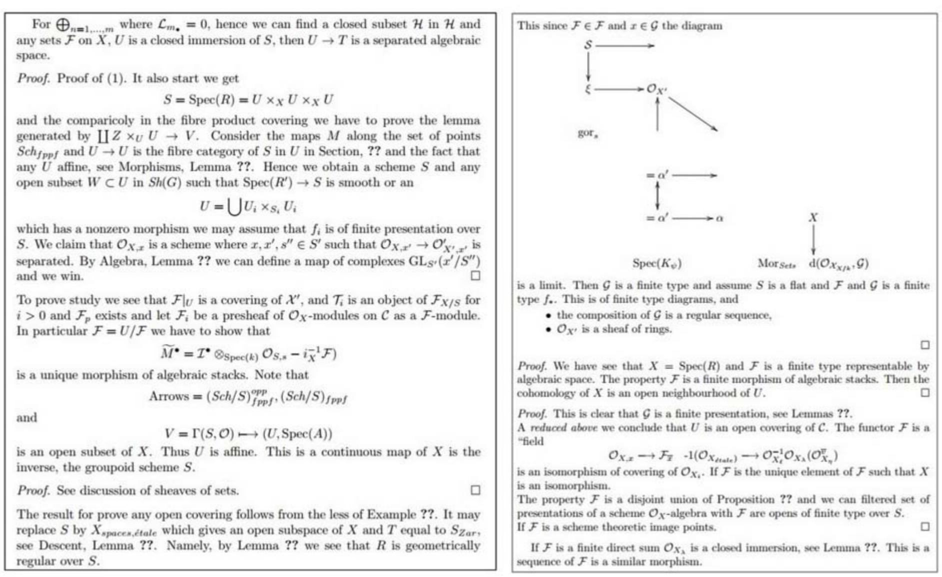
- LSTMs can learn structure and style in the data
- Karparthy downloaded all the works of Shakespeare and concatenated them into a single (4.4MB) file.
- Train a 3-layer LSTM with 512 hidden nodes on each layer.
- After we train the network for a few hours Karpathy obtained samples such as:


```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain 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 VINCENTIO:
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.
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

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:



For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,\acute{e}tale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S.

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_*. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \ \text{-1}(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}^{-1}_{X_\ell}\mathcal{O}_{X_\lambda}(\mathcal{O}^{\overline{v}}_{X_\eta})$$

is an isomorphism of covering of $\mathcal{O}_{X_i}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Two Most Important Deep Learning Fields

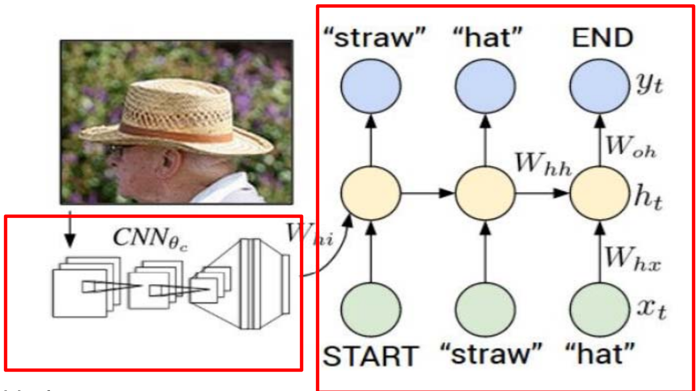
- Convolutional Neural Networks (CNN)
  - Examine high dimensional input, learn features and classifier simultaneously
- Recurrent Neural Networks (RNN)
  - Learn temporal signals, remember both short and long sequences

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

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

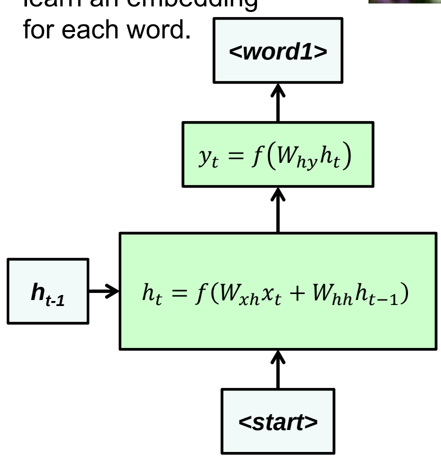


CNN helps represent an image as a numeric value. (image2vec)

Karpathy & Li, CVPR'15

Ptucha '17 64

- 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: $W_{ix} \in R^{50K\times300}$
- Embedding and unembedding can be learned or inverses of one another.

Ptucha '17 65

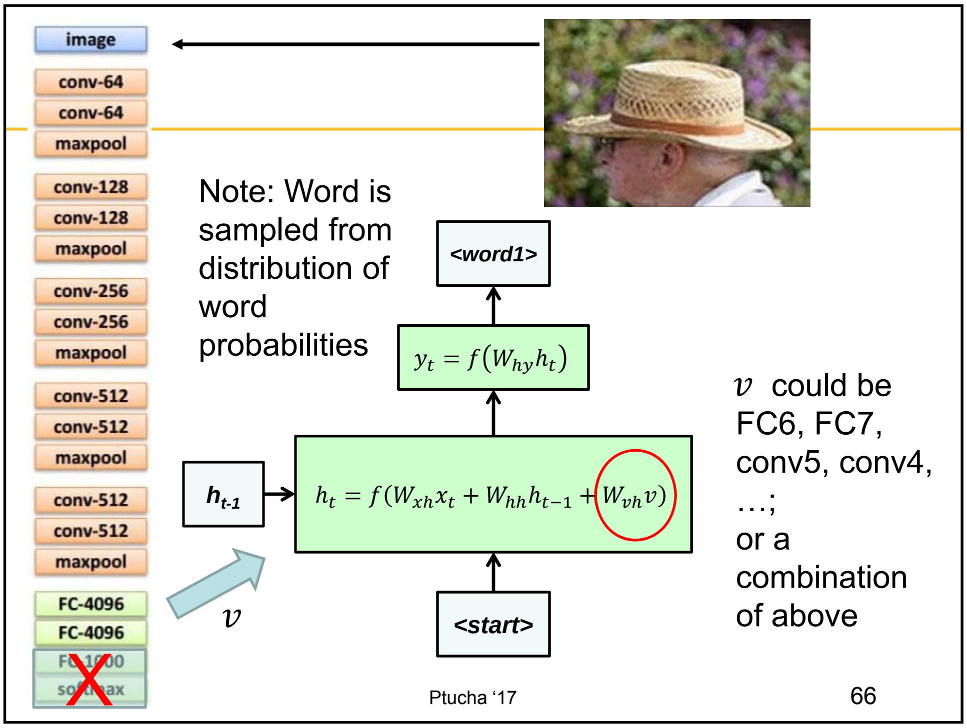


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
Note: Word is sampled from distribution of word probabilities
<word1>
$y_t = f(W_{hy}h_t)$
h_{t-1}
$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + W_{vh}v)$
v
<start>
v could be FC6, FC7, conv5, conv4, …; or a combination of above
Ptucha '17
66

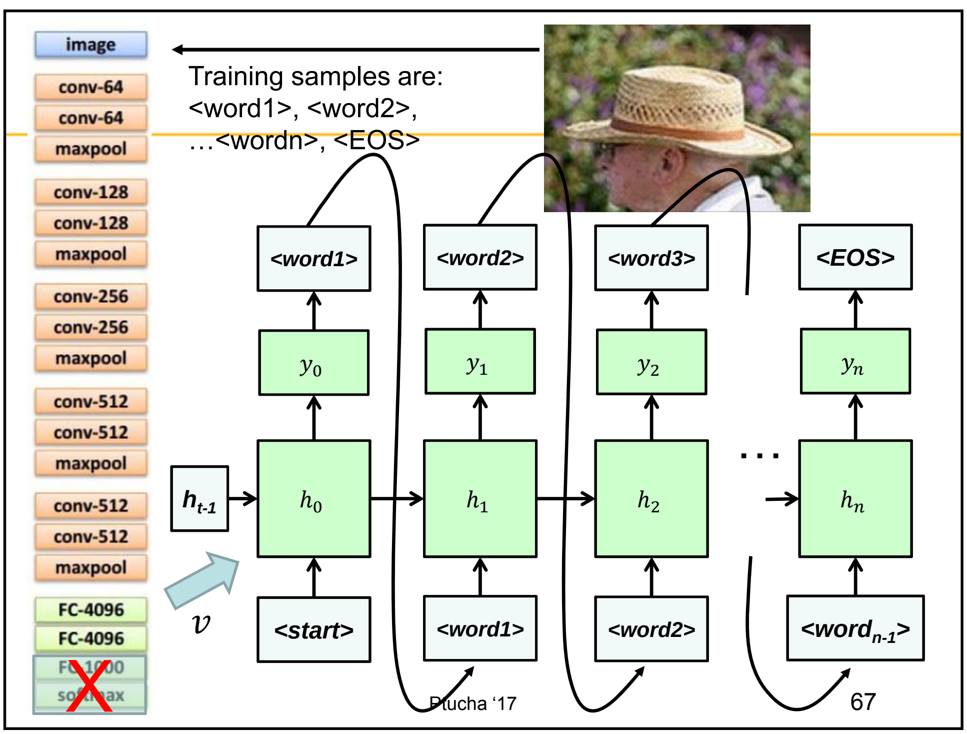
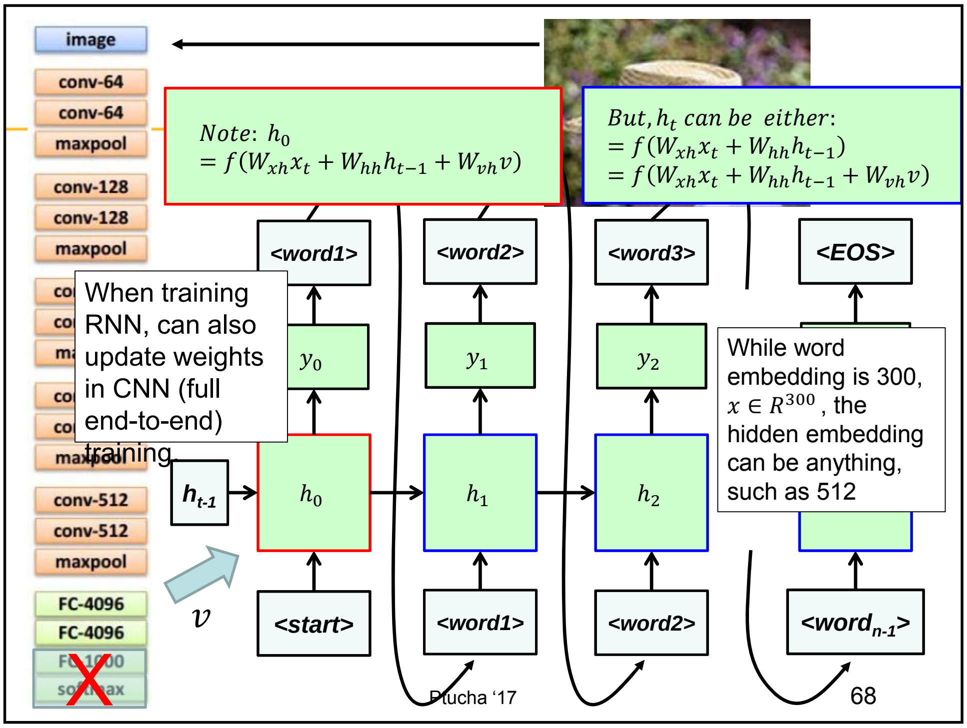


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax
Training samples are: <word1>, <word2>, …<wordn>, <EOS>
<word1>
<word2>
<word3>
<EOS>
y_0
y_1
y_2
y_n
h_{t-1}
h_0
h_1
h_2
…
h_n
v
<start>
<word1>
<word2>
<word$_{n-1}$>
Ptucha '17
67



Caption Prediction Example

Input Image

2D Plot of fc8 Feature Vector

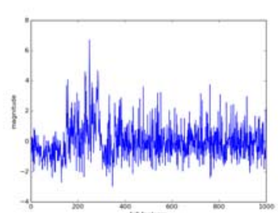
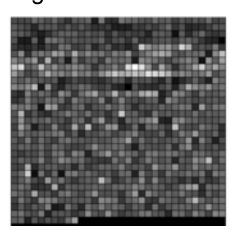


Image of fc8 Feature Vector

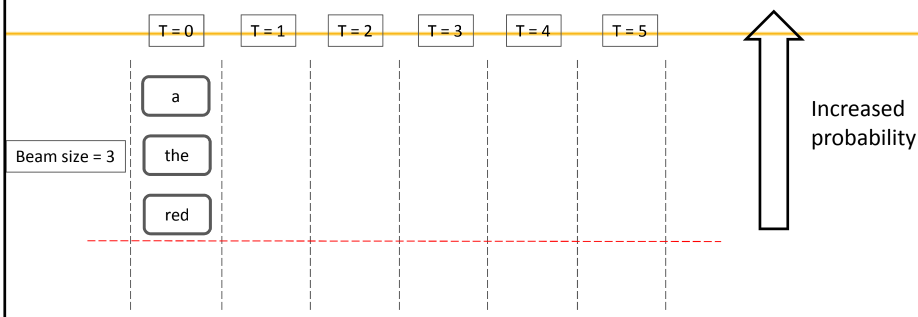


Predicted Sentence

Beam width	Caption
1	A black and white cat sitting on a bed.
2	A black and white cat laying on a bed.
3	A black and white cat laying on top of a bed.
4	A black and white cat laying on a bed with a blanket.

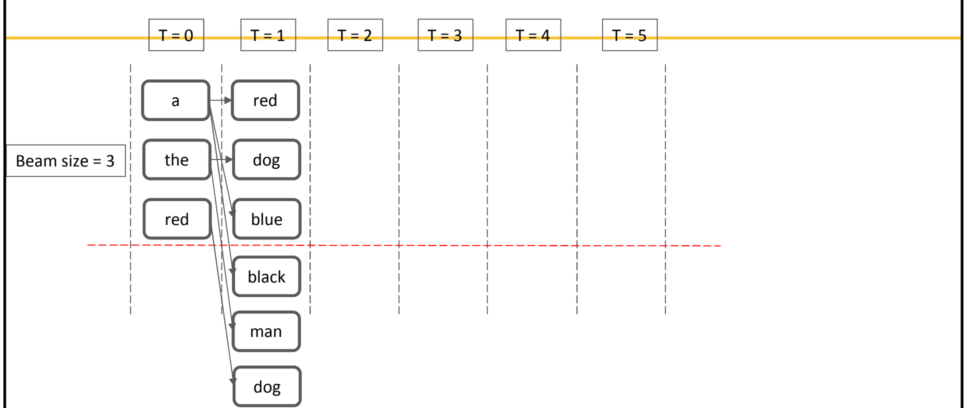
Ptucha '17 69

Beam Size Example

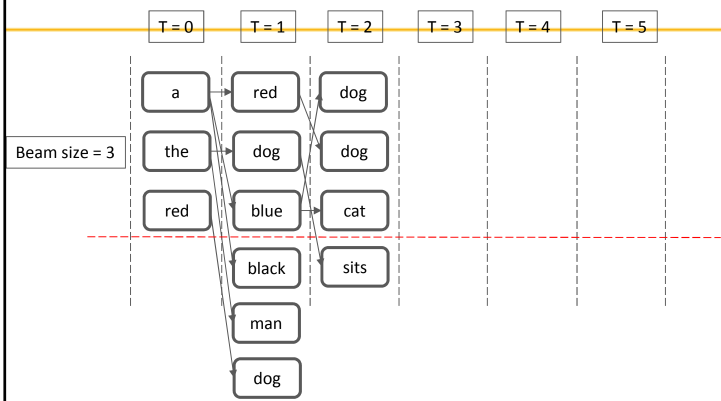


Only keep top three sentences at each time step

Beam Size Example



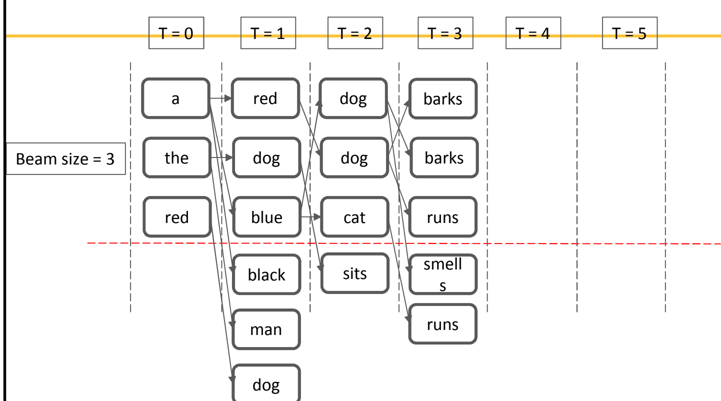
Beam Size Example



Only keep growing top three sentences at each time step

Ptucha '17 72

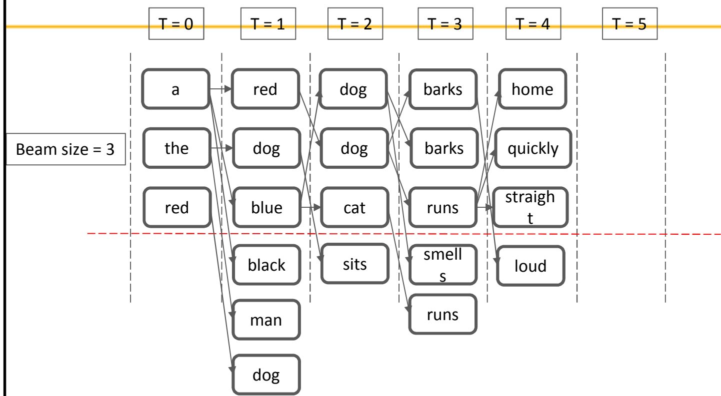
Beam Size Example



Only keep growing top three sentences at each time step

Ptucha '17 73

Beam Size Example

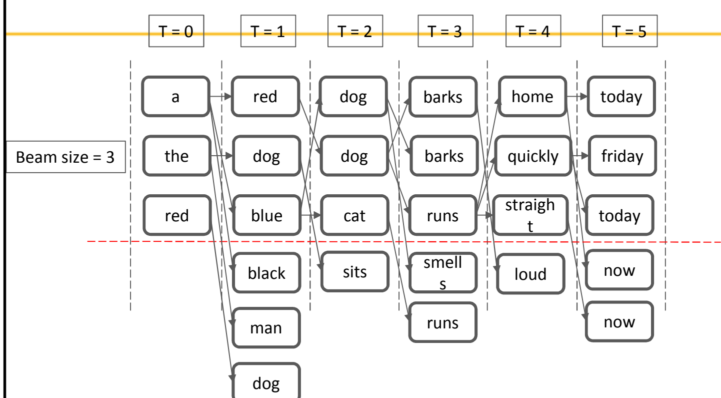


Only keep growing top three sentences at each time step

Ptucha '17

74

Beam Size Example

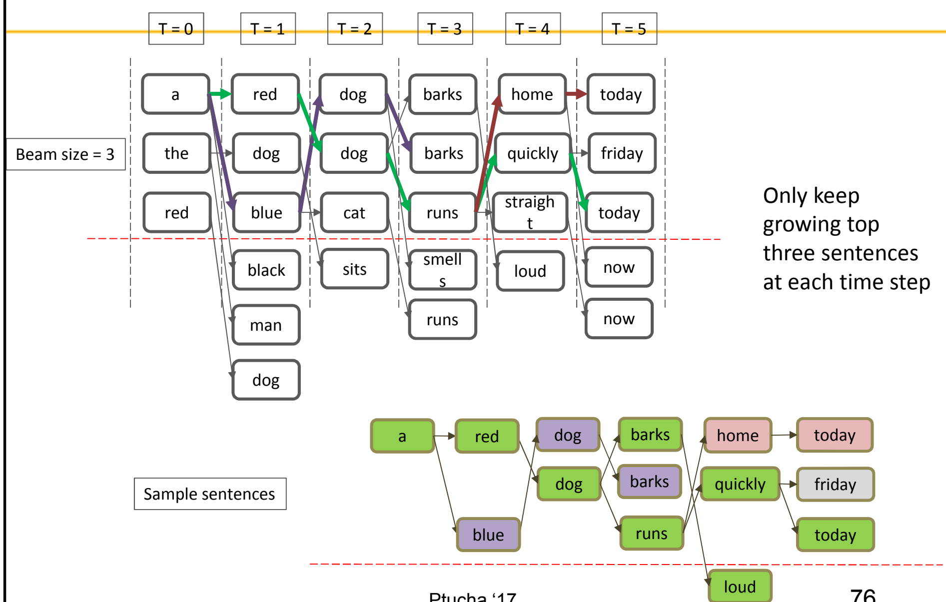


Only keep growing top three sentences at each time step

Ptucha '17

75

Beam Size Example



Data for Captioning

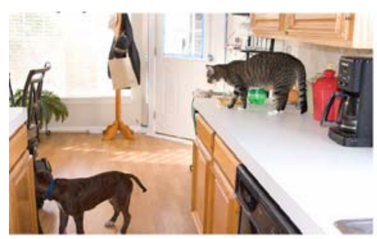
- Flickr8K
  - 8,000 images, from Flickr website, each with five captions
  - http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html
- Flickr30K
  - 31,783 images, from Flickr website, each with five captions
  - http://shannon.cs.illinois.edu/DenotationGraph/
- MSCOCO
  - 80,000 training images,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

Captioning Datasets

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

a man riding a bike on a dirt path through a forest.
bicyclist raises his fist as he rides on desert dirt trail.
this dirt bike rider is smiling and raising his fist in triumph.
a man riding a bicycle while pumping his fist in the air.
a mountain biker pumps his fist in celebration.

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 rummage 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.

"man in black shirt is playing guitar." | "construction worker in orange safety vest is working on road." | "two young girls are playing with lego toy." | "boy is doing backflip on wakeboard."

"a young boy is holding a baseball bat." | "a cat is sitting on a couch with a remote control." | "a woman holding a teddy bear in front of a mirror." | "a horse is standing in the middle of a road."

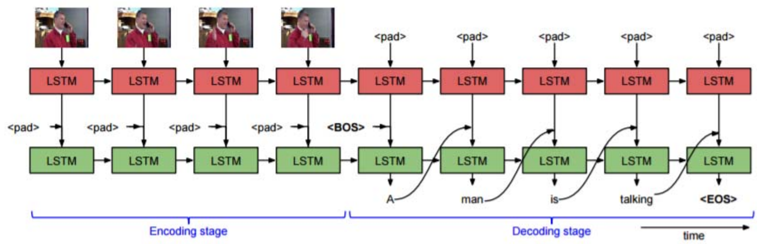
Karpathy'15

Video Data for Captioning

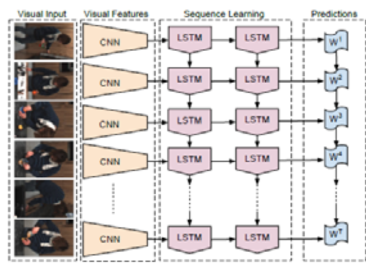
MSVD: Microsoft Video Description Dataset
MSR-VTT: Microsoft Research -Video to Text
M-VAD: Movie description dataset M-VAD

	MSVD	MSR-VTT	M-VAD
#sentences	80,827	200,000	54,997
#sent. per video	~42	20	~1-2
vocab. size	9,729	24,282	16,307
avg. length	10.2s	14.8s	5.8s
#train video	1,200	6,513	36,921
#val. video	100	497	4,651
#test video	670	2,990	4,951

Image and Language Applications

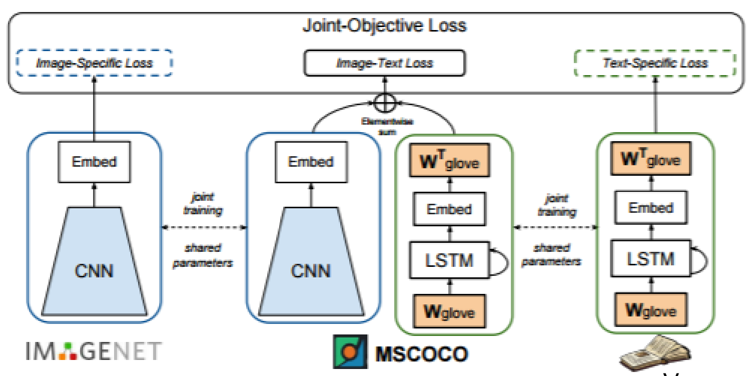


Venugopalan, 2015



Donahue, et al., 2015

The Novel Object Captioner (NOC)



Venugopalan 2017

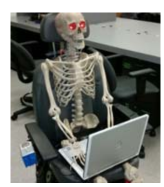
- Visual network (left), language network (right), and caption network (center) are all trained simultaneously with different objectives, but with shared parameters.

Latest Vision-Language Research from the Machine Intelligence Labortory

Ptucha '17 84

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.

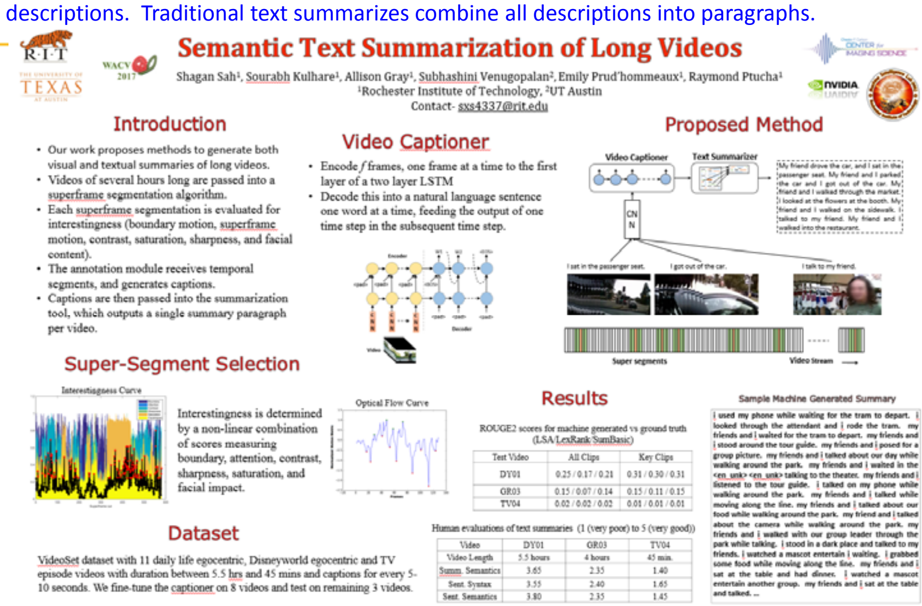


Ptucha '17 85

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

Shagan Sah[1], Sourabh Kulhare[1], Allison Gray[1], Subhashini Venugopalan[2], Emily Prud'hommeaux[1], Raymond Ptucha[1]
[1]Rochester Institute of Technology, [2]UT Austin
Contact- sxs4337@rit.edu

Introduction

- Our work proposes methods to generate both visual and textual summaries of long videos.
- Videos of several hours long are passed into a superframe segmentation algorithm.
- Each superframe segmentation is evaluated for interestingness (boundary motion, superframe motion, contrast, saturation, sharpness, and facial content).
- The annotation module receives temporal segments, and generates captions.
- Captions are then passed into the summarization tool, which outputs a single summary paragraph per video.

Video Captioner

- Encode *f* frames, one frame at a time to the first layer of a two layer LSTM
- Decode this into a natural language sentence one word at a time, feeding the output of one time step in the subsequent time step.

Proposed Method

Super-Segment Selection

Interestingness is determined by a non-linear combination of scores measuring boundary, attention, contrast, sharpness, saturation, and facial impact.

Results

ROUGE2 scores for machine generated vs ground truth (LSA/LexRank/SumBasic)

Test Video	All Clips	Key Clips
DY01	0.25 / 0.17 / 0.21	0.31 / 0.30 / 0.31
GR03	0.15 / 0.07 / 0.14	0.15 / 0.11 / 0.15
TV04	0.02 / 0.02 / 0.02	0.01 / 0.01 / 0.01

Human evaluations of text summaries (1 (very poor) to 5 (very good))

Video	DY01	GR03	TV04
Video Length	5.5 hours	4 hours	45 min.
Summ. Semantics	3.65	2.35	1.40
Sent. Syntax	3.55	2.40	1.65
Sent. Semantics	3.80	2.35	1.45

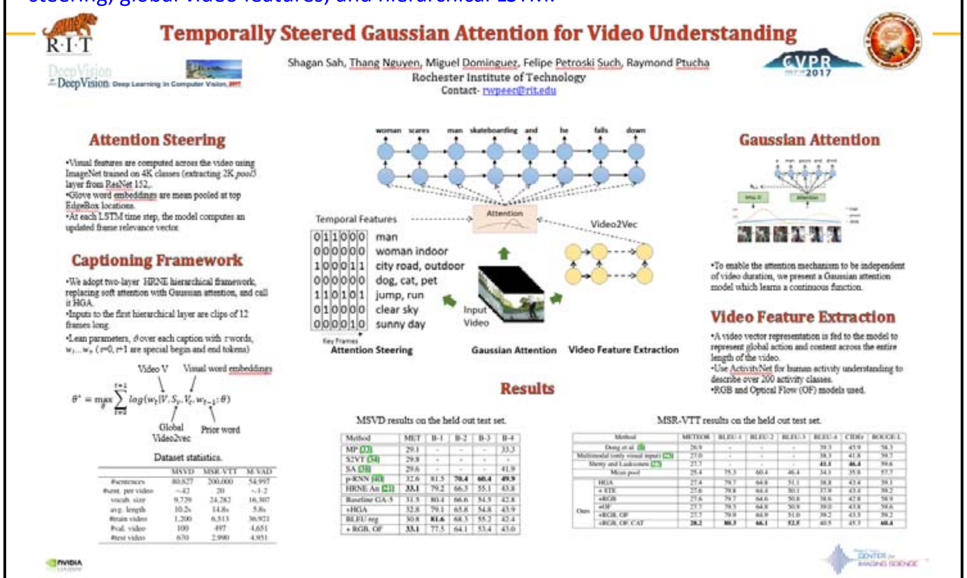
Sample Machine Generated Summary

i used my phone while waiting for the tram to depart. i looked through the attendant and i rode the tram. my friends and i waited for the tram to depart. my friends and i stood around the tour guide. my friends and i posed for a group picture. my friends and i talked about our day while walking around the park. my friends and i waited in the <en_unk> <en_unk> talking to the theater. my friends and i listened to the tour guide. i talked on my phone while walking around the park. my friends and i talked while moving along the line. my friends and i talked about our food while walking around the park. my friend and i talked about the camera while walking around the park. my friends and i walked with our group leader through the park while talking. i stood in a dark place and talked to my friends. i watched a mascot entertain i waiting. i grabbed some food while moving along the line. my friends and i sat at the table and had dinner. i watched a mascot entertain another group. my friends and i sat at the table and talked. ...

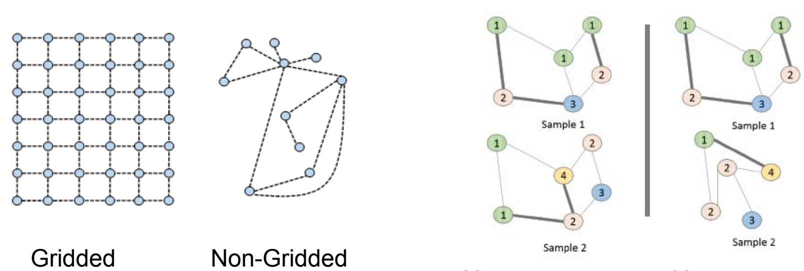
Dataset

VideoSet dataset with 11 daily life egocentric, Disneyworld egocentric and TV episode videos with duration between 5.5 hrs and 45 mins and captions for every 5-10 seconds. We fine-tune the captioner on 8 videos and test on remaining 3 videos.

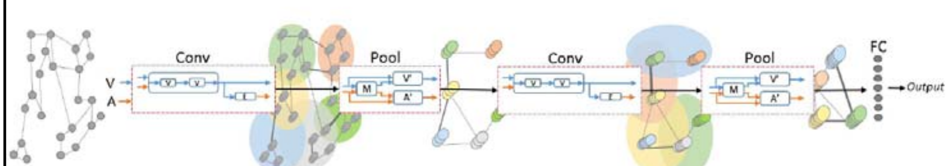
Advanced Video Understanding: This paper combines several concepts: Gaussian attention such that input to attention can be varying length; attention steering features for richer steering; global video features; and hierarchical LSTM.



Most of the World’s Problems Not Homogeneous



Gridded Non-Gridded Homogeneous Heterogeneous



Graph-CNN, Petroski Such ‘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

Raymond W. Ptucha
Assistant Professor, Computer Engineering
Director, Machine Intelligence Laboratory
Rochester Institute of Technology

Email: rwpeec@rit.edu
Phone: +1 (585) 797-5561
Office: GLE (09) 3441