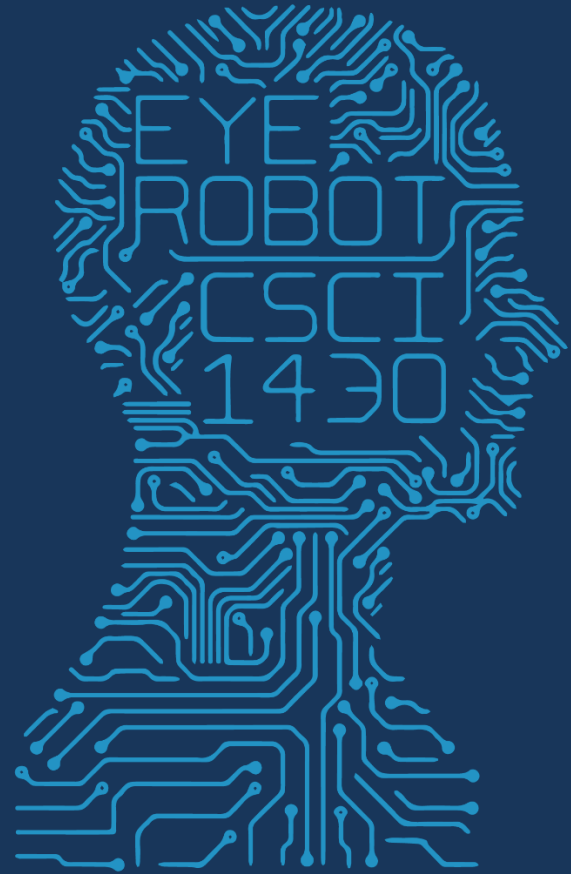




1950

FUTURE VISION



2017 MWF 1PM 368

COMPUTER VISION

# We're going to read real papers

- I haven't read them.
  - Puts me in the same position as you - let's try and work them out!
  - Not lazy, *honest*.
- Today: image captioning
  - Accessibility for visually impaired

## SHARE

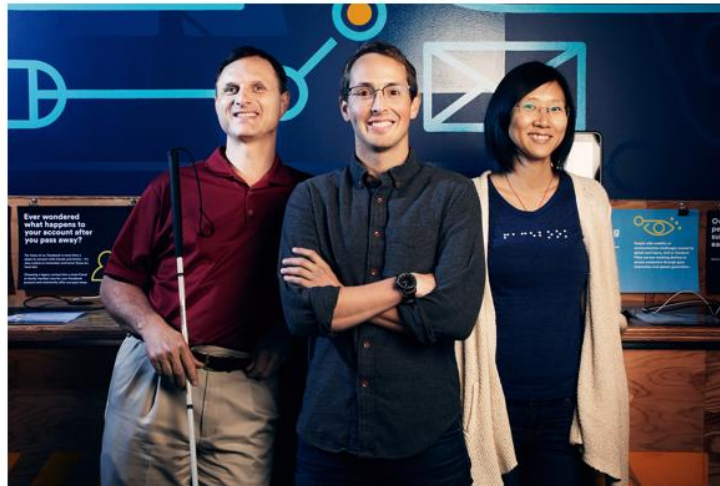
SHARE  
1613

TWEET

COMMENT  
0

EMAIL

CADE METZ BUSINESS 10.27.15 9:00 AM

FACEBOOK'S AI CAN CAPTION  
PHOTOS FOR THE BLIND ON ITS  
OWN

Facebook's Matt King, Jeff Wieland, and Shaomei Wu. FACEBOOK

MATT KING IS blind, so he can't see the photo. And though it was posted to his Facebook feed with a rather lengthy caption, that's no help. Thanks to text-to-speech software, his laptop reads the caption aloud, but it's in German. And King doesn't understand German.

But then he runs an artificial intelligence tool under development at Facebook, and after analyzing the photo, the tool goes a long way towards describing it. The scene is outdoors, the AI says. It includes grass and trees and clouds.

## MOST POPULAR



INTERNET CULTURE  
Inside the Conspiracy  
Theory That Turned  
Syria's First Responders  
EMMA GREY ELLIS



GEAR  
Our 10 Favorite Laptops,  
From MacBooks to  
Chromebooks  
WIRED STAFF



SCANDALS  
Red Light Cameras May Be  
Teeming Sama Tinkate

## SHARE

SHARE  
506

TWEET



EMAIL

CADE METZ BUSINESS 04.05.16 12:01 AM

FACEBOOK'S AI IS NOW  
AUTOMATICALLY WRITING  
PHOTO CAPTIONS

3 / 3 FACEBOOK

Facebook is now using artificial intelligence to automatically generate captions for photos in the News Feed of people who can't see them.

The tool is called Automatic Alternative Text, and it dovetails

## MOST POPULAR



GEAR  
Our 10 Favorite Laptops,  
From MacBooks to  
Chromebooks  
WIRED STAFF



INTERNET CULTURE  
Inside the Conspiracy  
Theory That Turned  
Syria's First Responders  
EMMA GREY ELLIS



SCANDALS  
Red Light Cameras May Be  
Issuing Some Tickets  
Based on Basic Math

# CVPR2015 – Deep Visual-Semantic Alignments for Generating Image Descriptions

- [Andrej Karpathy](#)
- Li Fei-Fei
- Department of Computer Science, Stanford University
- <http://cs.stanford.edu/people/karpathy/cvpr2015.pdf>
- <https://github.com/karpathy/neuraltalk2>

# Reading academic papers

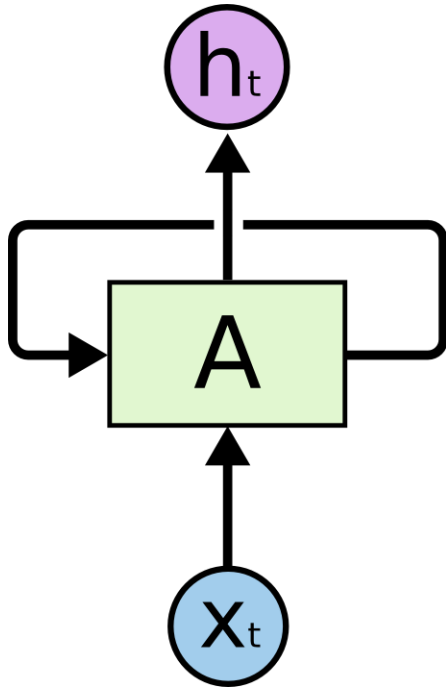
- 1<sup>st</sup> Pass:
  - 10 minutes
- Read title, abstract
- Look at section / subsection titles – not the paragraphs!
- Look at figures / tables to gain an overview.
- OK to skip things you don't quite understand.
- Answer:
  - What is the task? What are they trying to accomplish?
  - At a high level, how is this accomplished?
  - What is the outcome? How is this assessed?

# Reading academic papers

- 2<sup>nd</sup> pass
  - 20 minutes
- Start to look at details
- Read technical body
- Look at figures / tables in detail
- Look at related work, look up things you don't know
- Answer:
  - How does it actually work?
  - How is it different from existing works?
  - Does this seem like a reasonable approach? Limitations?

# What's an RNN?

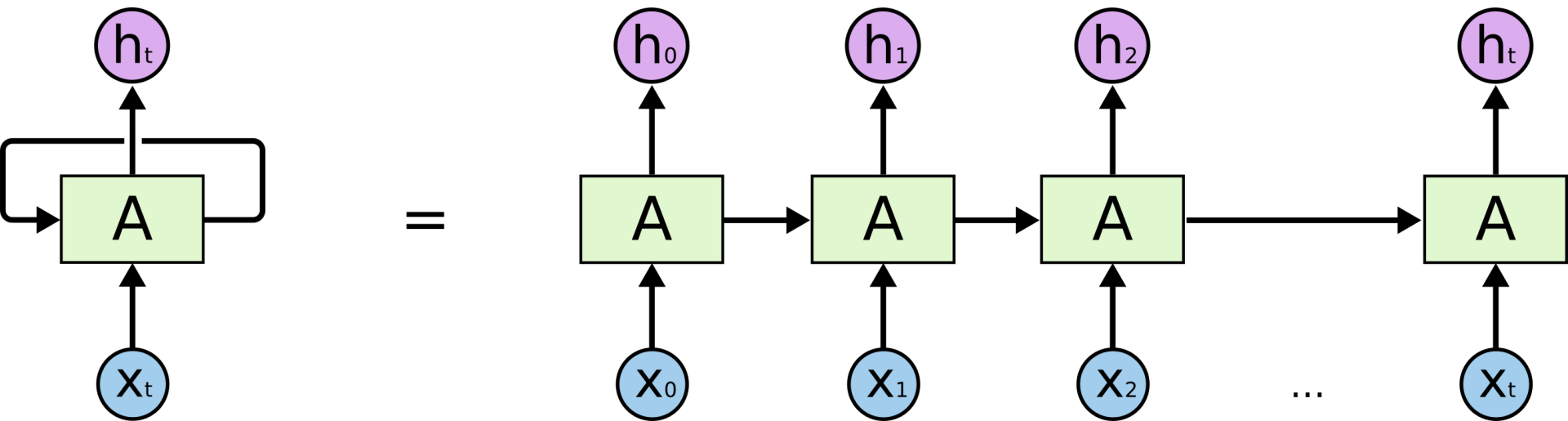
- Recurrent Neural Network
- Try to connect previous information to the present task.

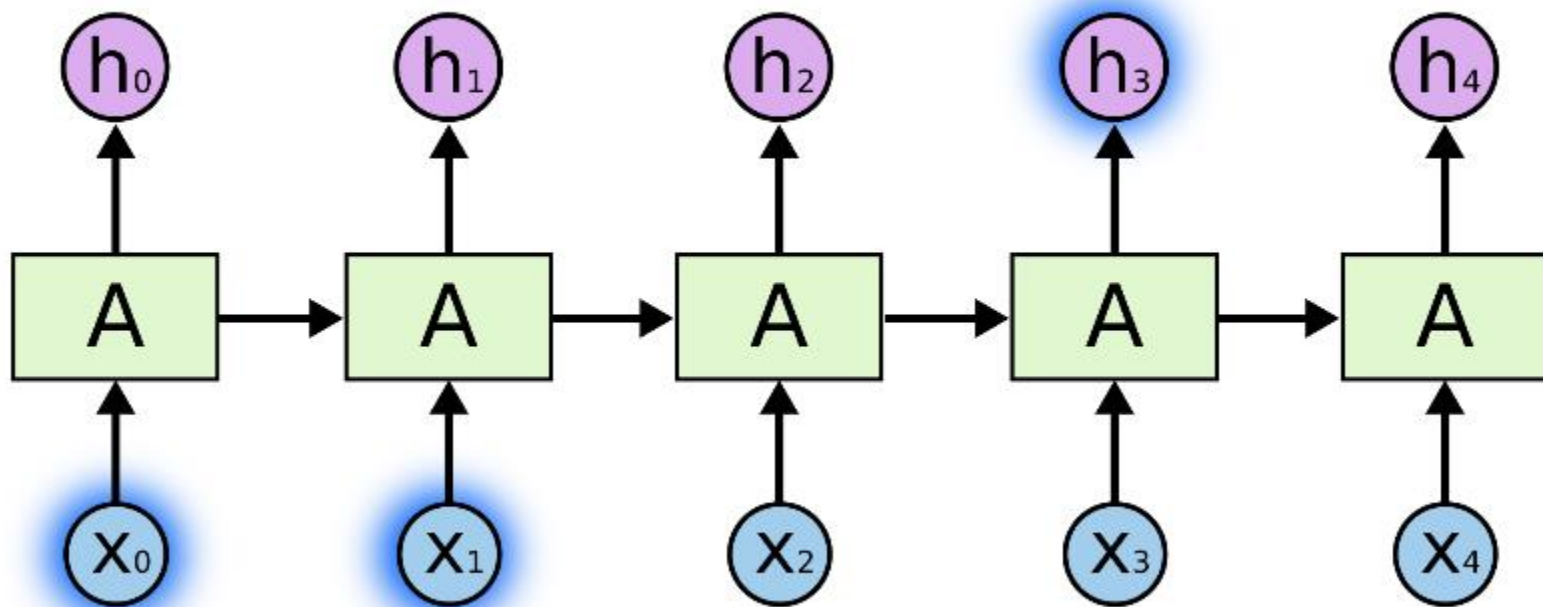


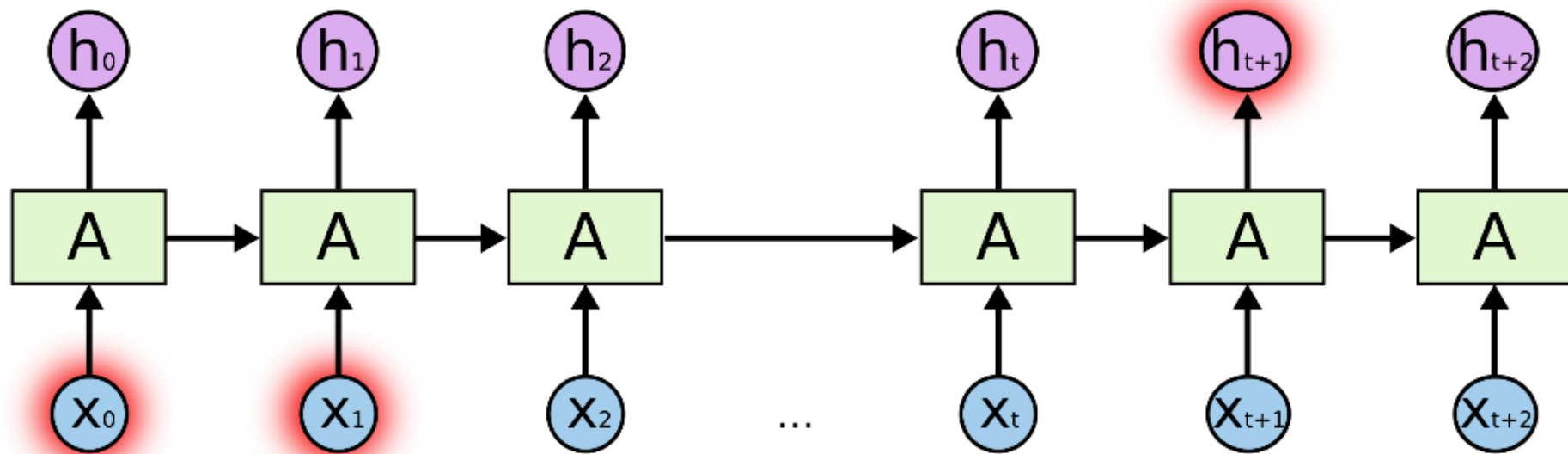


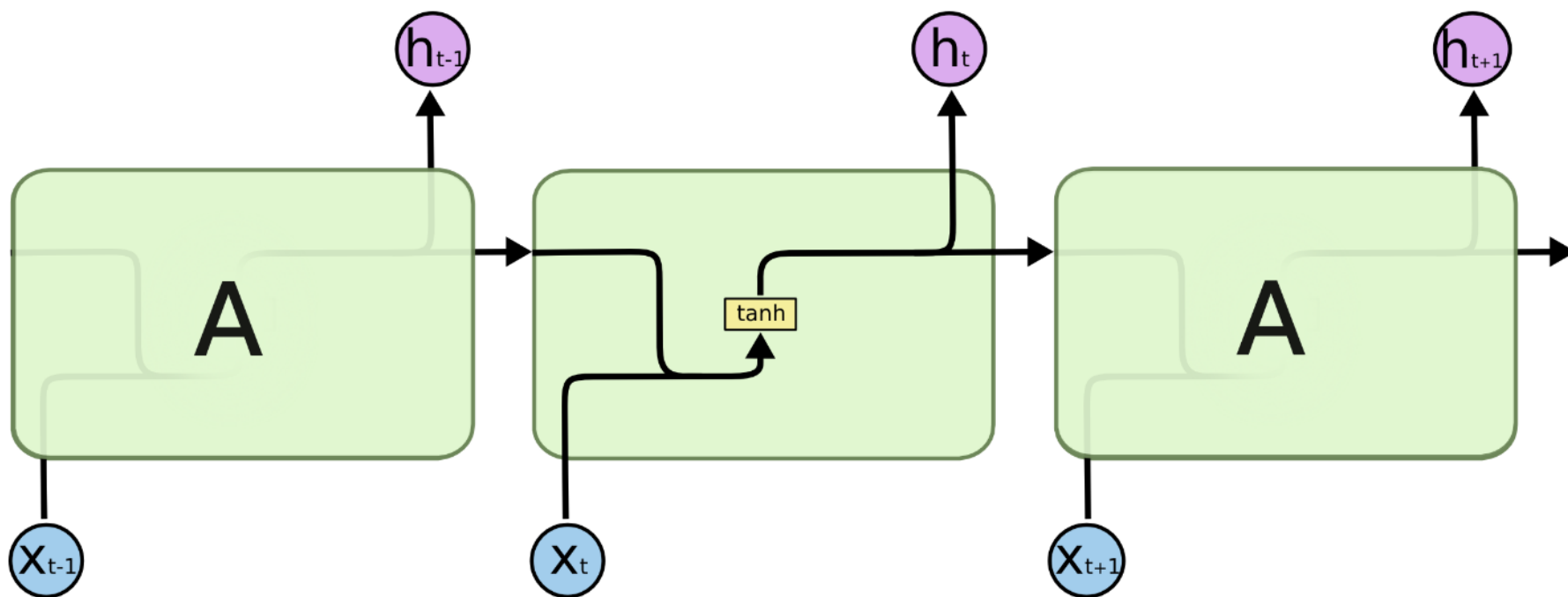
# What's an RNN?

- Recurrent Neural Network
- Try to connect previous information to the present task.



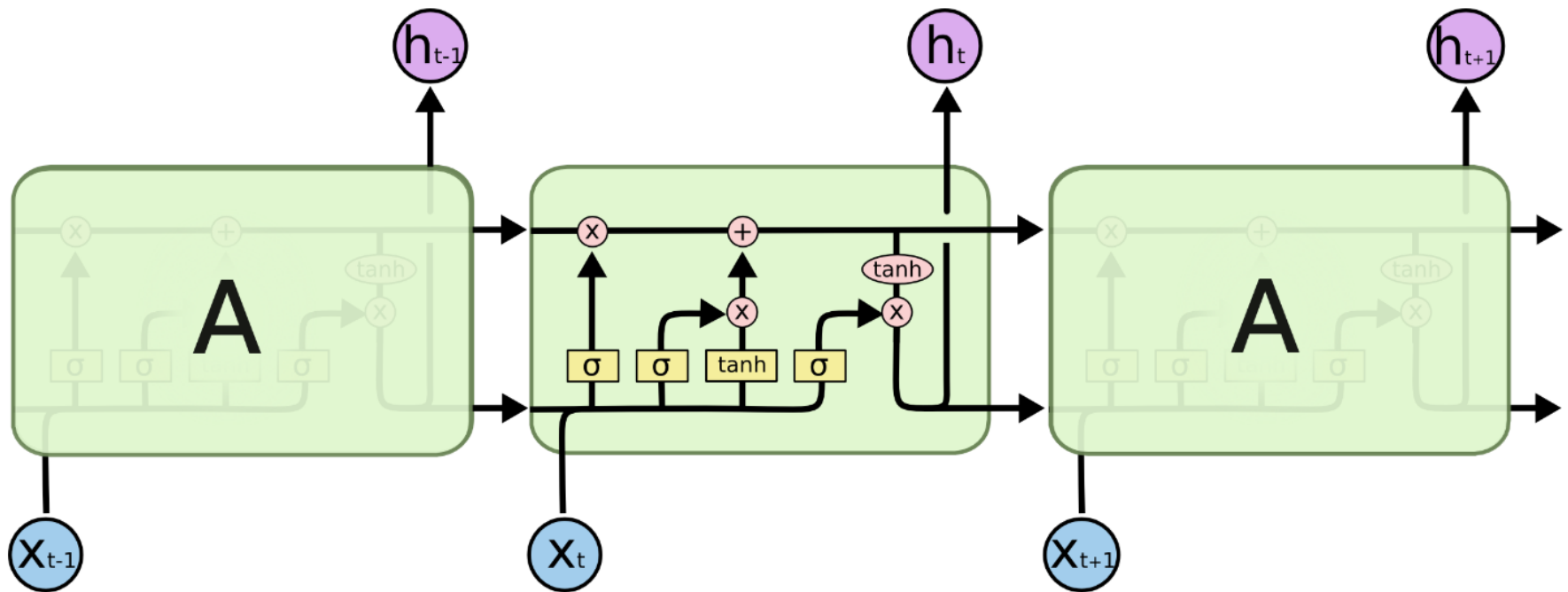






The repeating module in a standard RNN contains a single layer.

LSTM = Long Short Term Memory



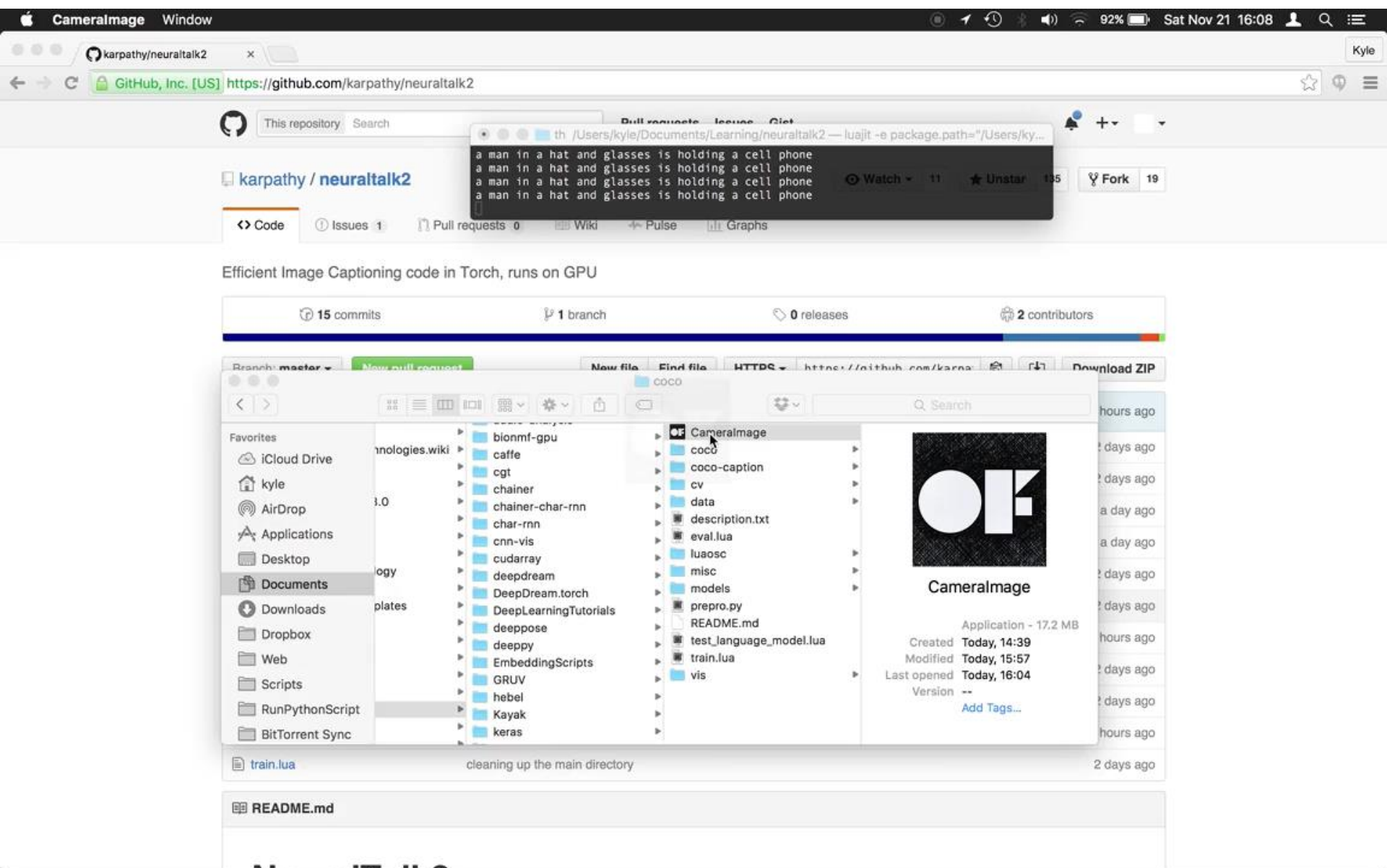
The repeating module in an LSTM contains four interacting layers.

# Reading academic papers

- 3<sup>rd</sup> pass
  - Not for today
- Mental re-implementation of the work
  - Start with the same assumptions as the authors
  - What would you do?
  - Compare your idea to the paper.
- Identify and challenge every assumption

## Answer:

- Strengths and weaknesses
- Problems with experimental or analytical techniques



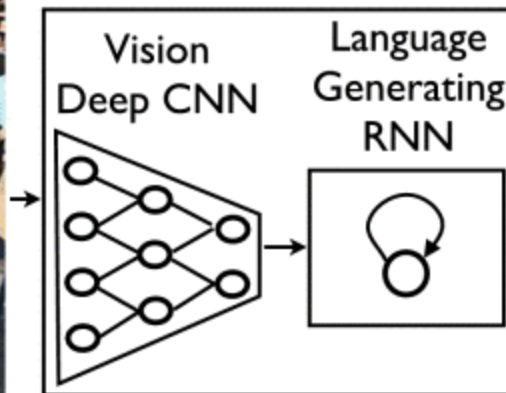
# Baidu Eye





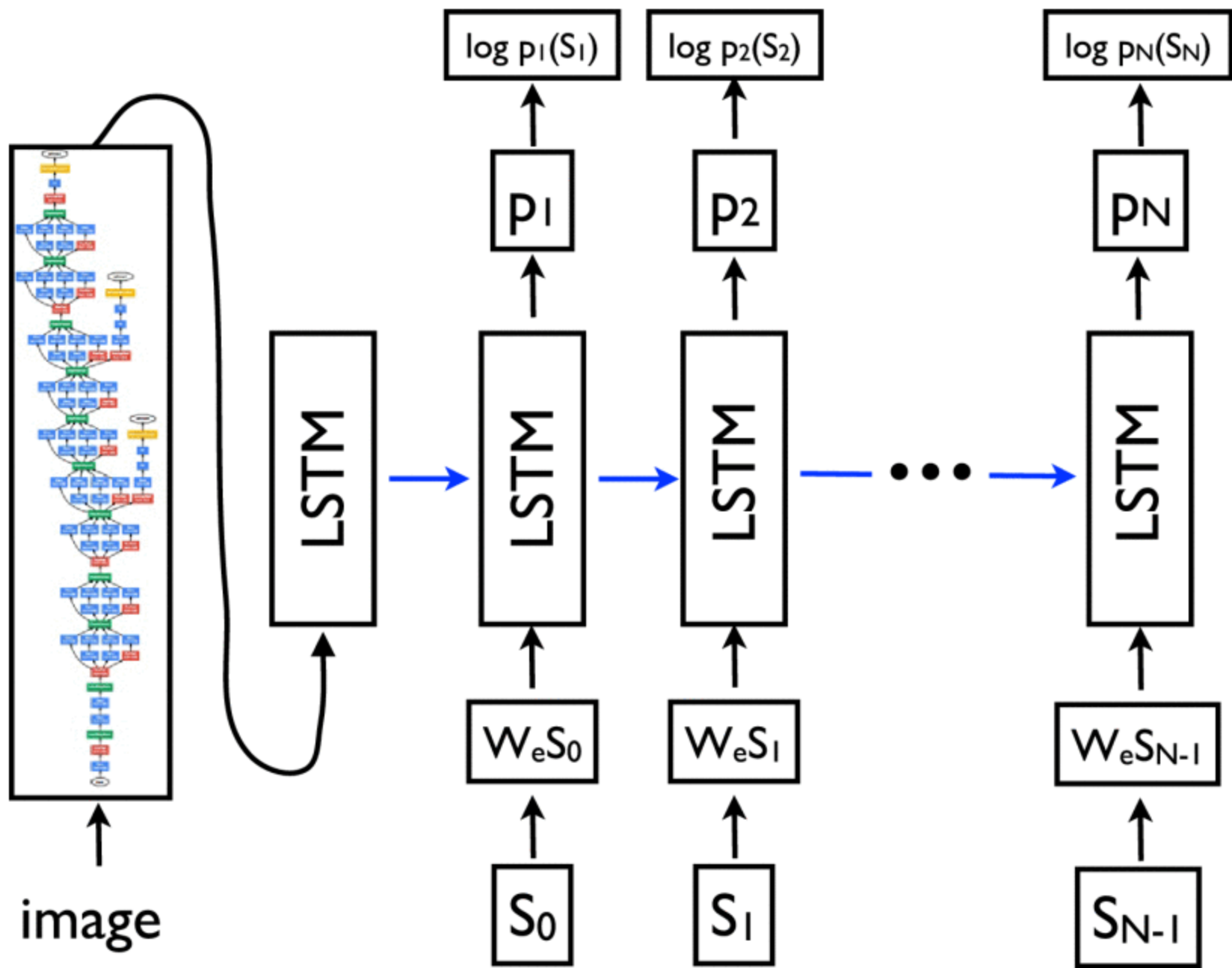
# TPAMI 2017: Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge

- Google version of the same thing
- <https://arxiv.org/abs/1609.06647>
- <http://ieeexplore.ieee.org/document/7505636/?arnumber=7505636>
- Implementation:
  - <https://github.com/tensorflow/models/tree/master/im2txt>
- “In our experience on an NVIDIA Tesla K20m GPU the initial training phase takes 1-2 weeks. The second training phase may take several additional weeks to achieve peak performance.”
  - ~5 days on a modern card.



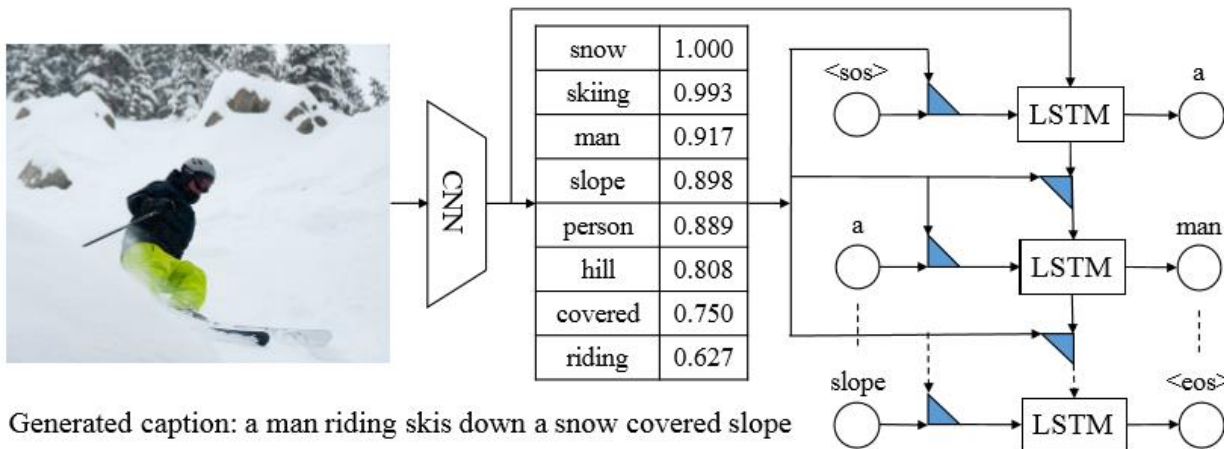
**A group of people  
shopping at an  
outdoor market.**

**There are many  
vegetables at the  
fruit stand.**




# CVPR 2017 - Semantic Compositional Networks for Visual Captioning

- [https://github.com/zhegan27/Semantic Compositional Nets](https://github.com/zhegan27/Semantic_Compositional_Nets)



(a) Overview of the proposed model.



**Detected semantic concepts:**  
 person (0.998), baby (0.983), holding (0.952), small (0.697), sitting (0.638), toothbrush (0.538), child (0.502), mouth (0.438)

**Semantic composition:**

1. Only using “**baby**”: *a baby in a*
2. Only using “**holding**”: *a person holding a hand*
3. Only using “**toothbrush**”: *a pair of toothbrush*
4. Only using “**mouth**”: *a man with a toothbrush*
5. Using “**baby**” and “**mouth**”: *a baby brushing its teeth*

**Overall caption generated by the SCN:**  
*a baby holding a toothbrush in its mouth*

**Influence the caption by changing the tag:**

6. Replace “**baby**” with “**girl**”: *a little girl holding a toothbrush in her mouth*
7. Replace “**toothbrush**” with “**baseball**”: *a baby holding a baseball bat in his hand*
8. Replace “**toothbrush**” with “**pizza**”: *a baby holding a piece of pizza in his mouth*

(b) Examples of SCN-based image captioning.

Human factors angle:

## CSCW 2017: Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service

- <https://research.fb.com/publications/automatic-alt-text-computer-generated-image-descriptions-for-blind-users-on-a-social-network-service/>
- Experiment with real users, measuring 'usefulness' of system

