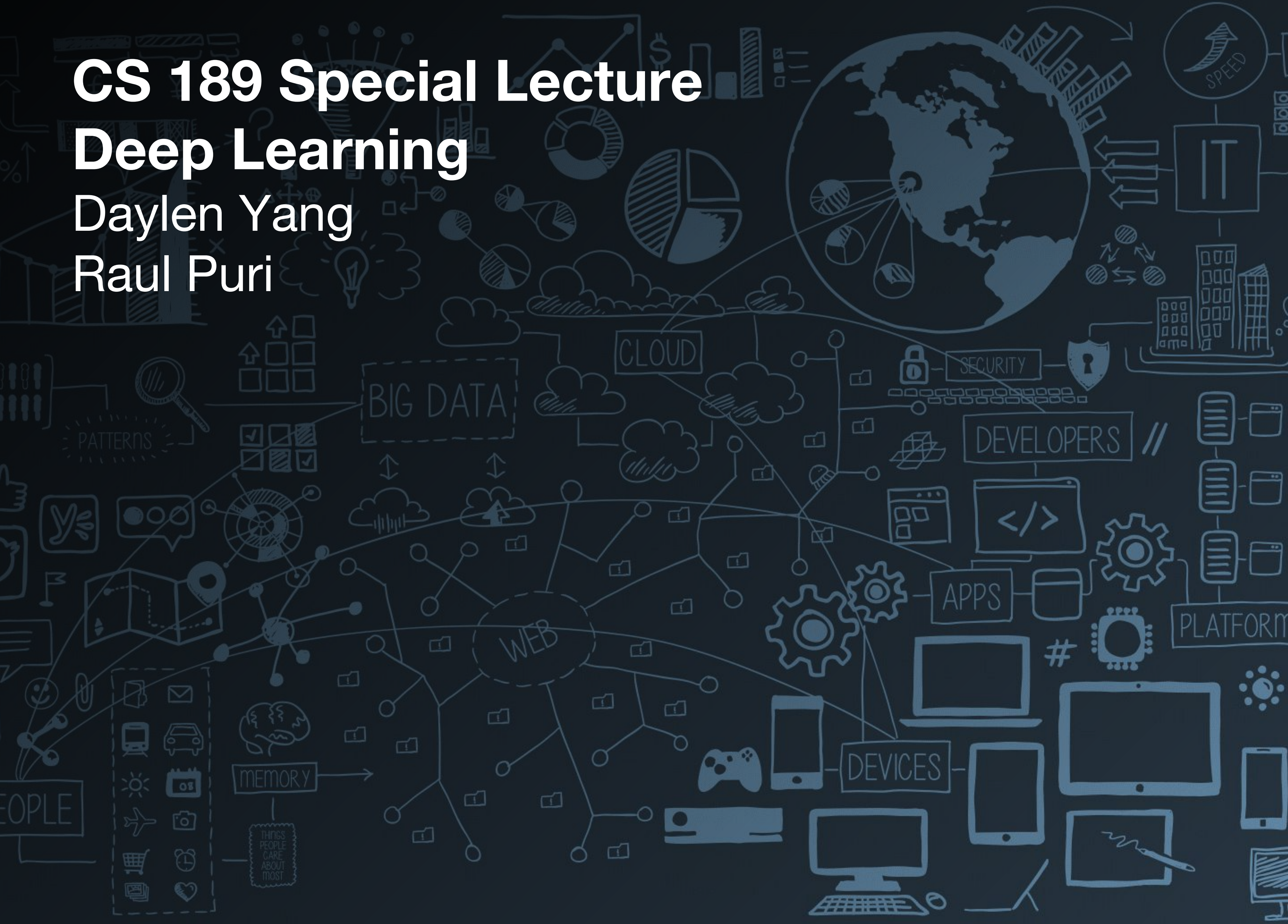
CS 189 Special Lecture
Deep Learning

Daylen Yang

Raul Puri



Logistics

- Second review session on Friday, 12-2pm, Wozniak Lounge
  - On post-midterm material
- Final exam is Monday, May 8, 3-6pm in RSF Fieldhouse

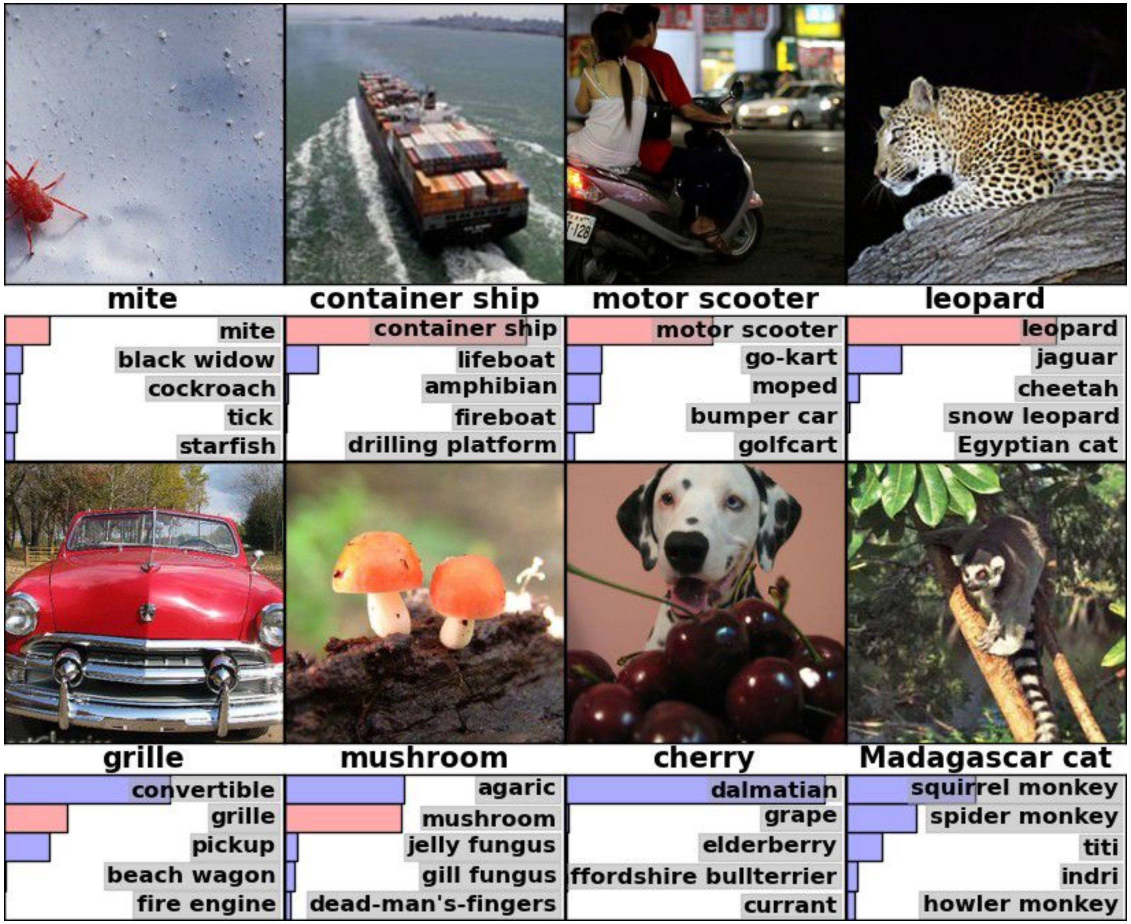
Today

- State-of-the-art ConvNets
- Recurrent Neural Networks
- Applications
    - Image Captioning
    - Visual Question Answering

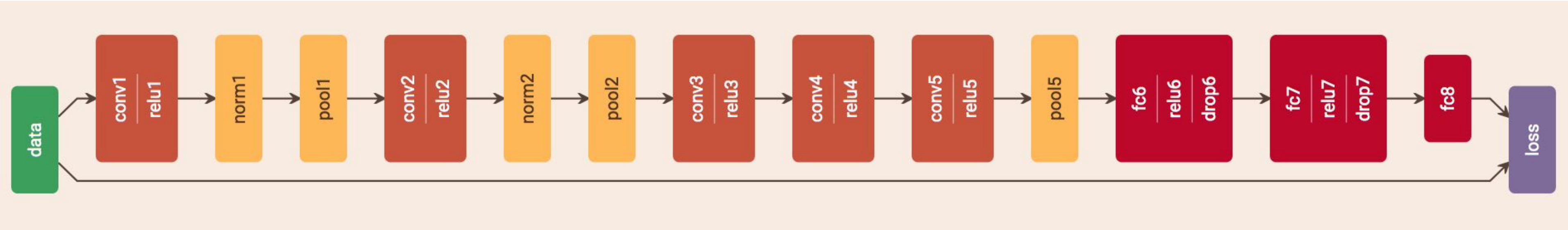
State-of-the-art ConvNet Architectures

AlexNet, GoogleNet, VGG, ResNet

The IMAGENET dataset



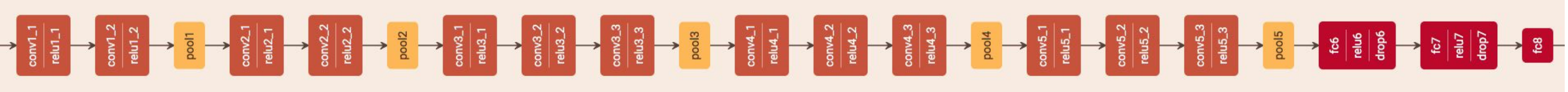
AlexNet (2012)



- 8 layers: 5 conv, 3 FC
- Top-5 error 16.4%
- Winner of ILSVRC 2012

http://ethereon.github.io/netscope

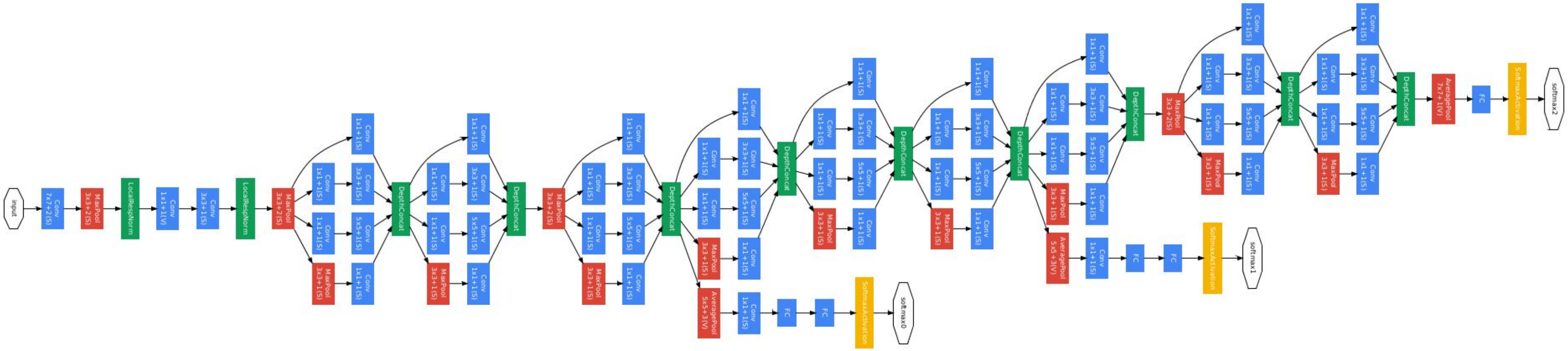
VGGNet (2014)



- 19 layers
- Top-5 error 7.3%
- 2nd place in ILSVRC 2014
- Notable for using 3x3 convolutions *everywhere*

http://ethereon.github.io/netscope

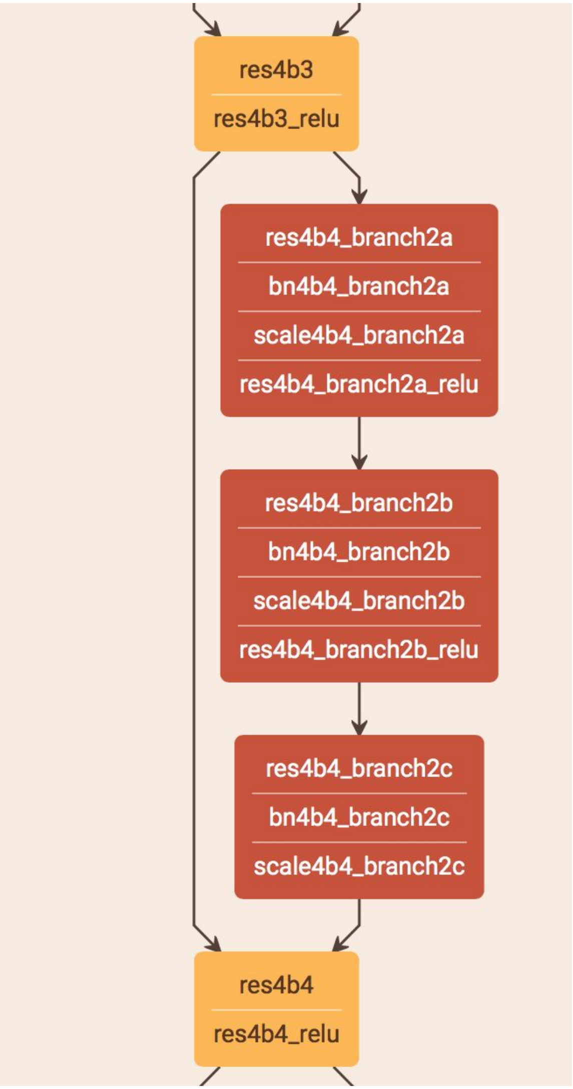
GoogleNet (2014)



- 22 layers
- Top-5 error 6.7%
- Winner of ILSVRC 2014
- Uses *Inception modules*

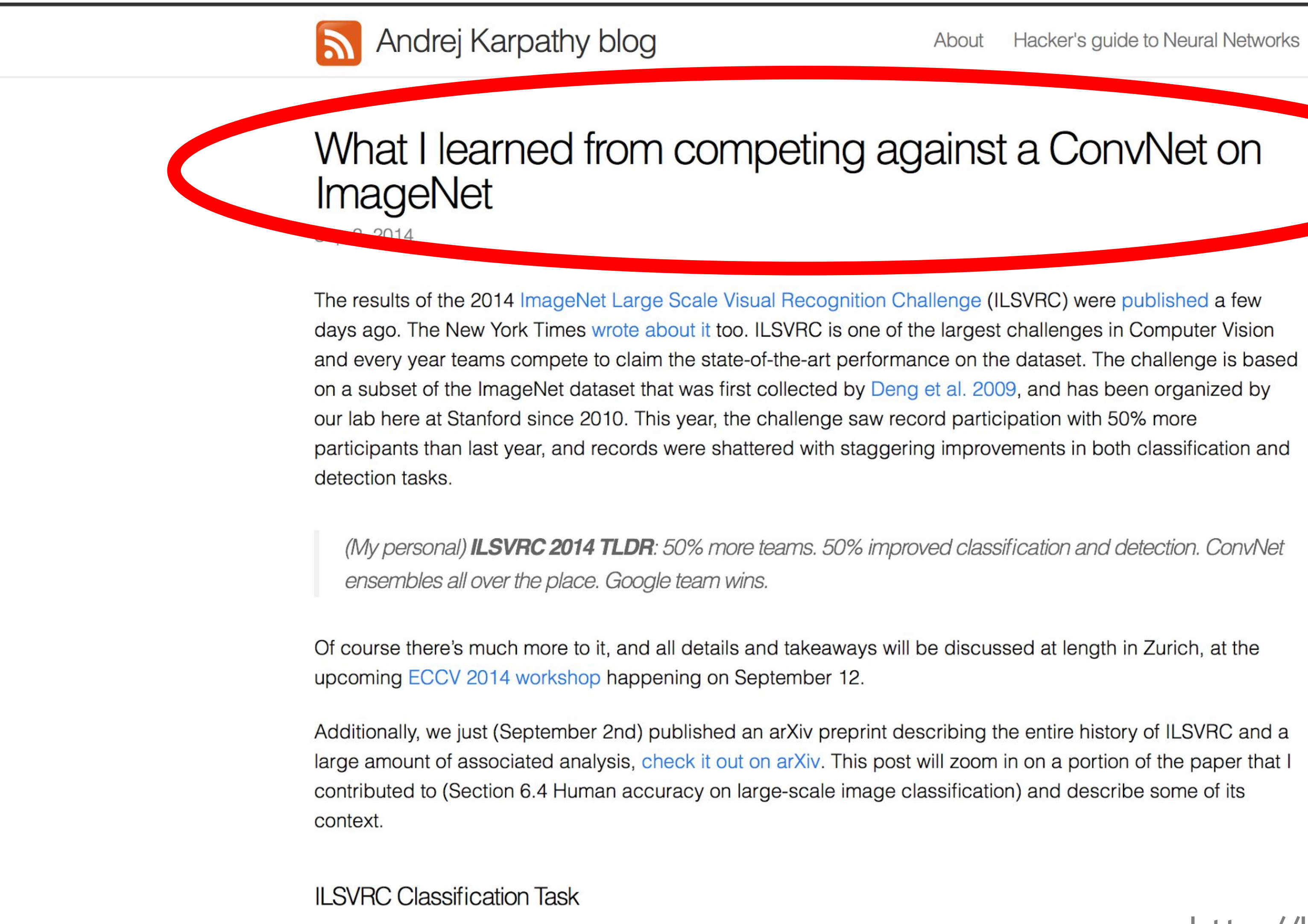
ResNet (2015)

- 152 layers (!)
- Top-5 error 3.57%
- Winner of ILSVRC 2015
- Uses skip connections to help gradient flow in deep neural networks



http://ethereon.github.io/netscope

Human Performance?



What I learned from competing against a ConvNet on ImageNet

Sep 2, 2014

The results of the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) were published a few days ago. The New York Times wrote about it too. ILSVRC is one of the largest challenges in Computer Vision and every year teams compete to claim the state-of-the-art performance on the dataset. The challenge is based on a subset of the ImageNet dataset that was first collected by Deng et al. 2009, and has been organized by our lab here at Stanford since 2010. This year, the challenge saw record participation with 50% more participants than last year, and records were shattered with staggering improvements in both classification and detection tasks.

> *(My personal)* ***ILSVRC 2014 TLDR****: 50% more teams. 50% improved classification and detection. ConvNet ensembles all over the place. Google team wins.*

Of course there's much more to it, and all details and takeaways will be discussed at length in Zurich, at the upcoming ECCV 2014 workshop happening on September 12.

Additionally, we just (September 2nd) published an arXiv preprint describing the entire history of ILSVRC and a large amount of associated analysis, check it out on arXiv. This post will zoom in on a portion of the paper that I contributed to (Section 6.4 Human accuracy on large-scale image classification) and describe some of its context.

ILSVRC Classification Task

5.1%

Top-5 error

http://karpathy.github.io

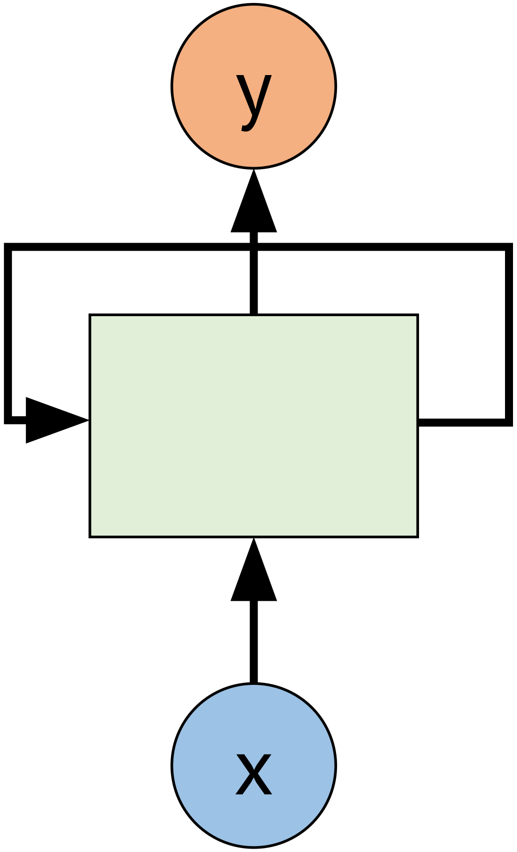
What’s Next in CV?

- Object detection
- Segmentation
- Video
- etc.

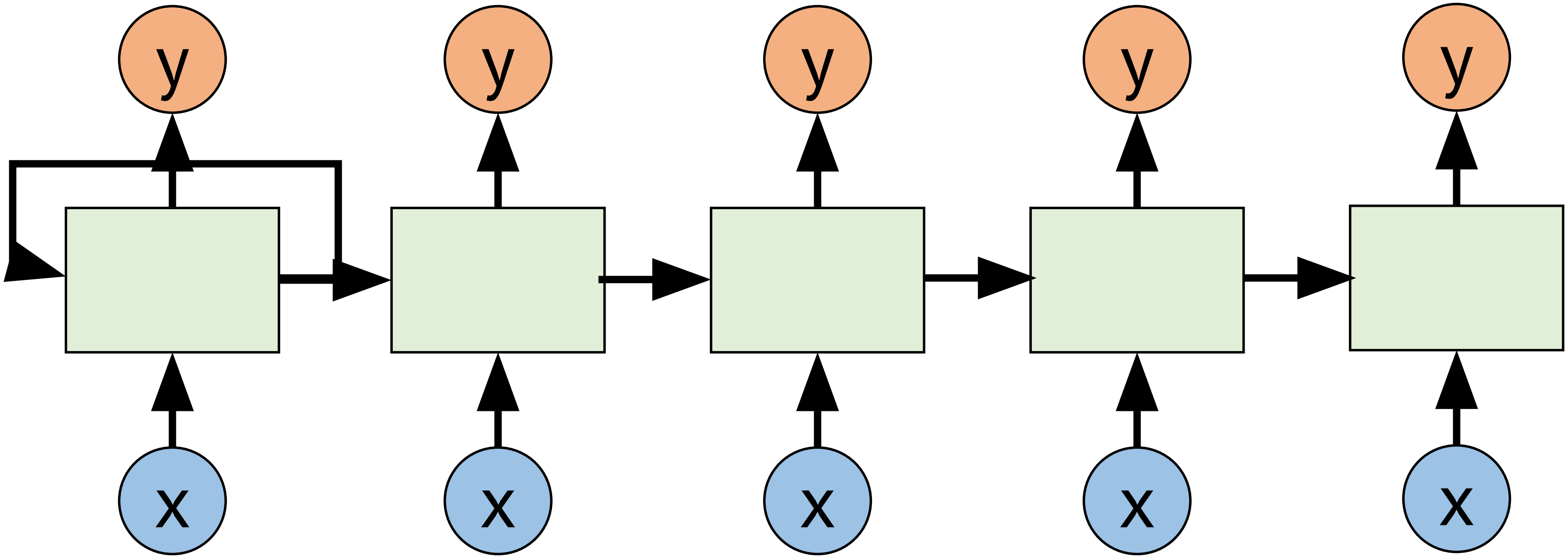
Recurrent Neural Networks

Why RNNs?

- To accept variable sized input!
- Applications:
  - Speech to text
  - Machine translation
  - Video classification
  - Attention mechanisms



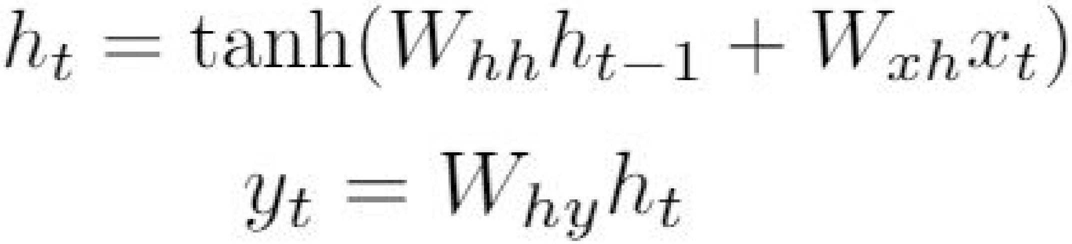
Unrolling a Recurrent Neural Network



RNN Unit

- Inputs: x and the previous hidden state
- Outputs: the next hidden state and y

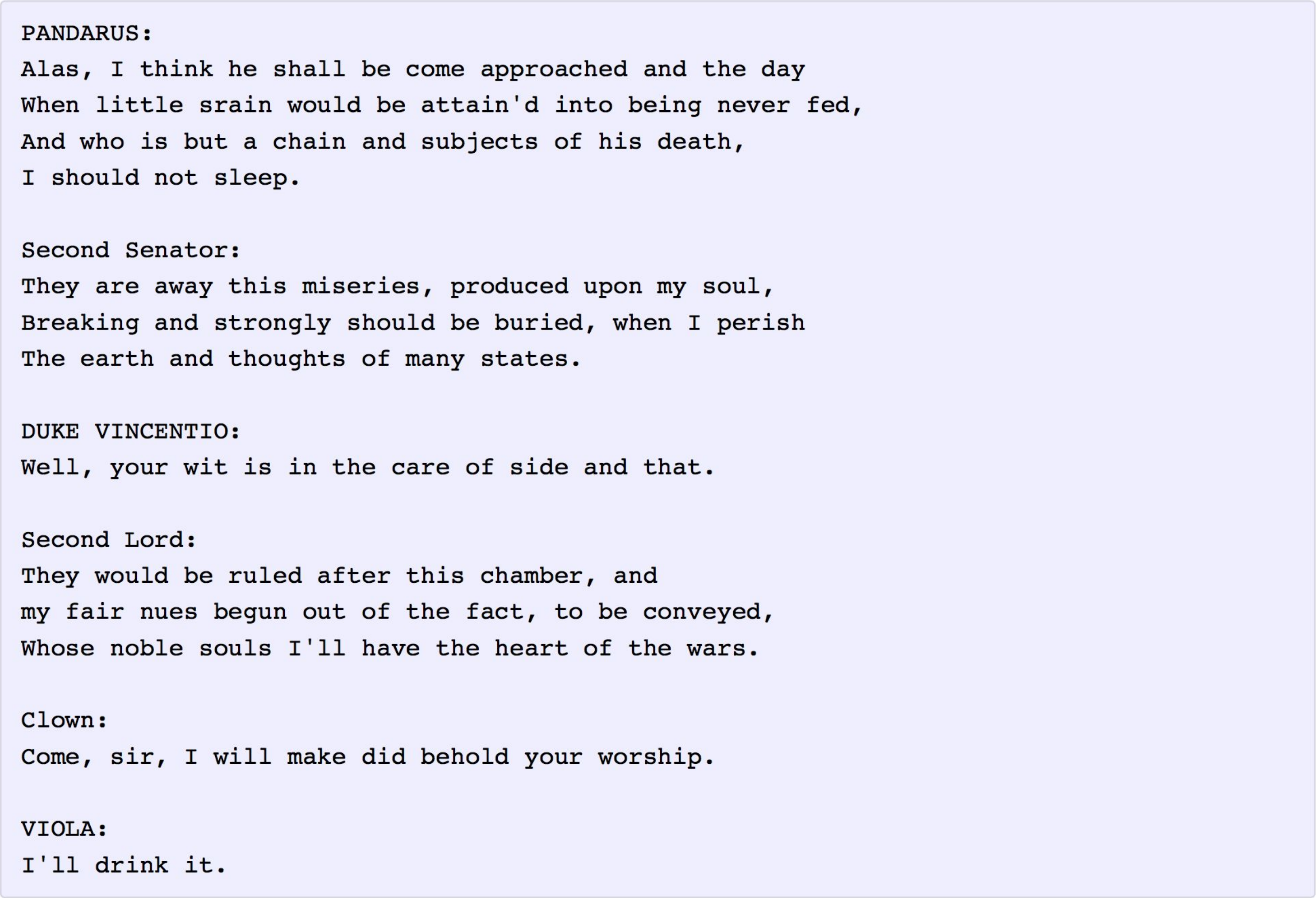
Forward Pass



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

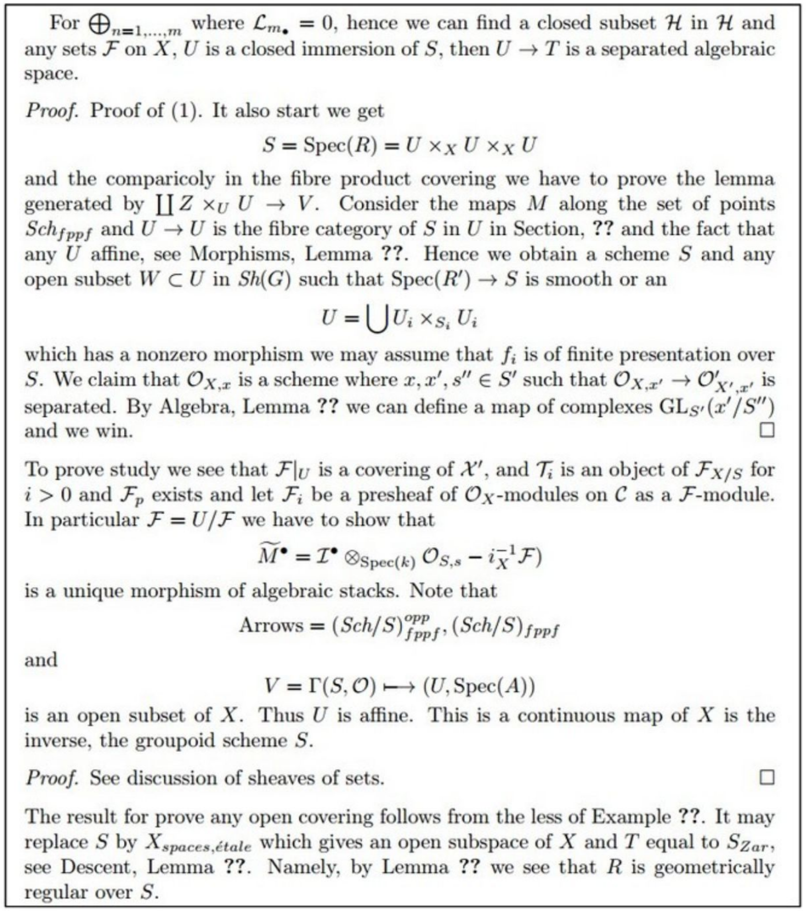
Generating Shakespeare



```
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.

VIOLA:
I'll drink it.
```

http://karpathy.github.io

Generating Latex



For $\bigoplus_{n=1,\dots,m}$ 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

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

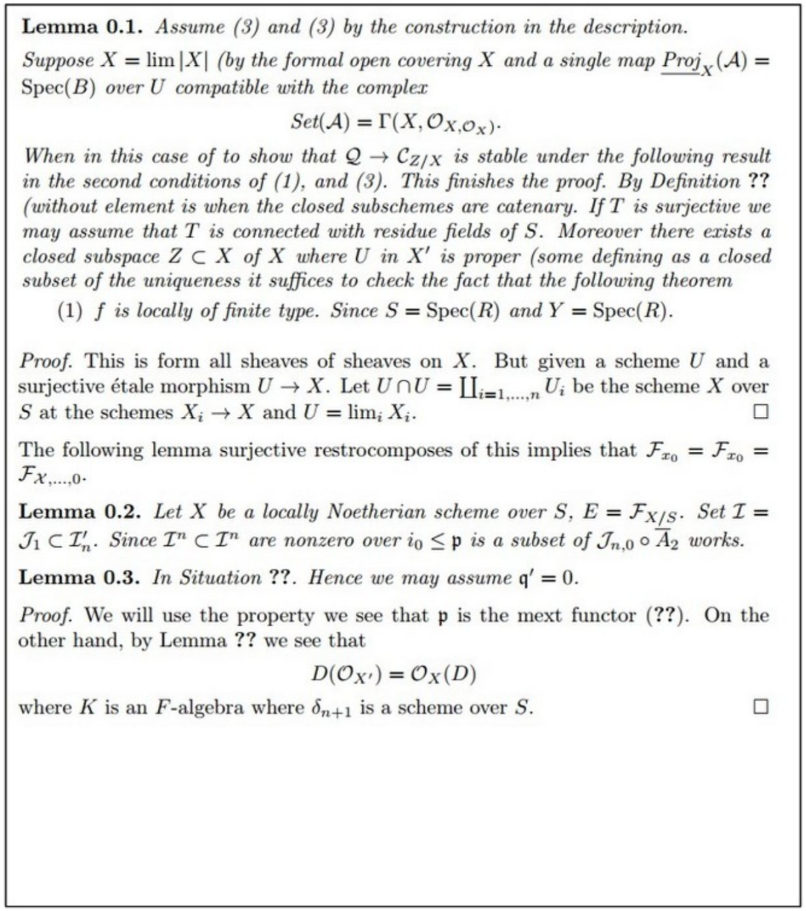
and

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

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 $X_{spaces,étale}$ 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.



Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition* **??** *(without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S*. Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\dots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\dots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S, $E = \mathcal{F}_{X/S}$*. Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$*. Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation* **??***. Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (**??**). On the other hand, by Lemma **??** we see that

$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

http://karpathy.github.io

Image Captioning

Demo

https://www.captionbot.ai

The Task

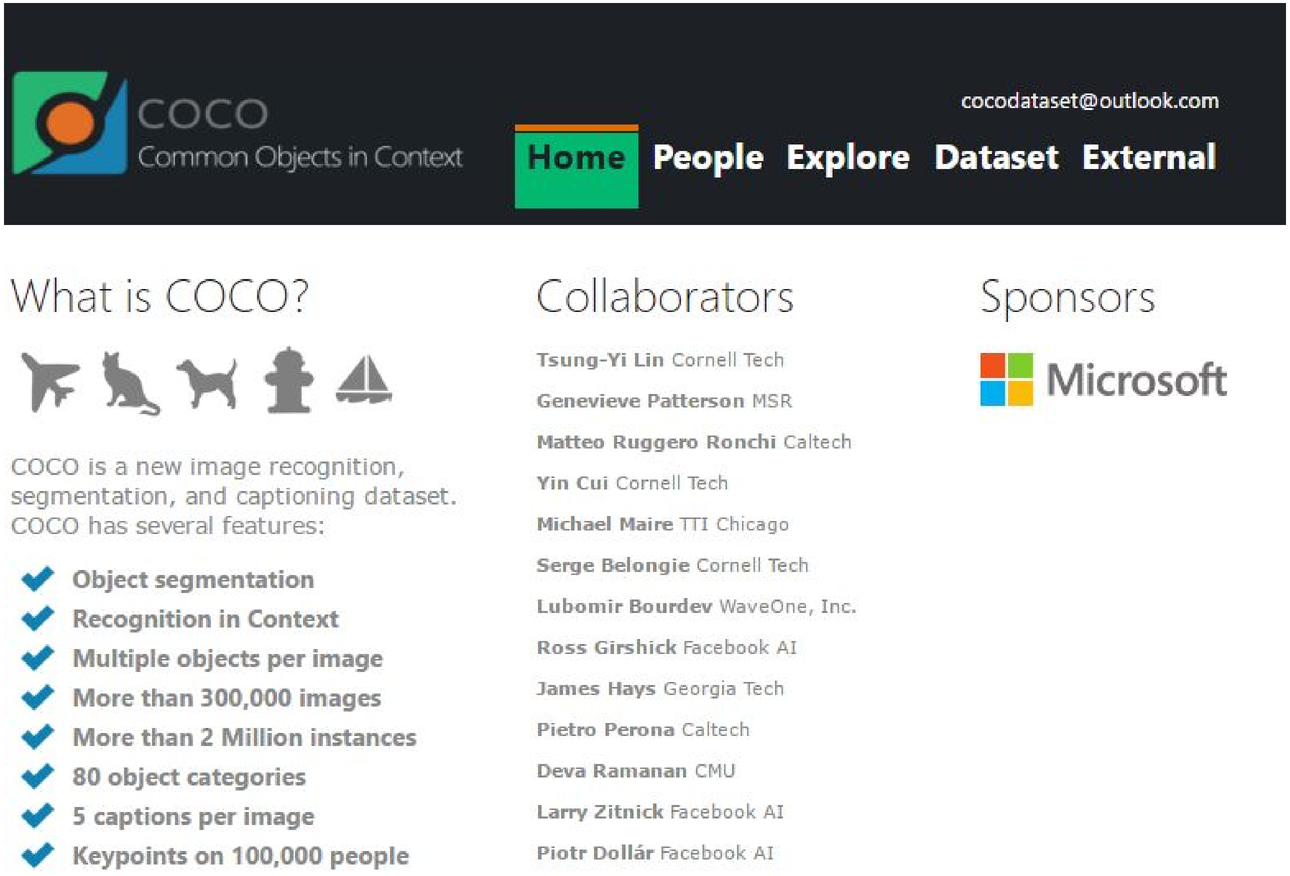
a man is playing tennis on a tennis court

a train is traveling down the tracks at a train station

a cake with a slice cut out of it

a bench sitting on a patch of grass next to a sidewalk

The Datasets - MS COCO



What is COCO?

COCO is a new image recognition, segmentation, and captioning dataset. COCO has several features:

- ✓ Object segmentation
- ✓ Recognition in Context
- ✓ Multiple objects per image
- ✓ More than 300,000 images
- ✓ More than 2 Million instances
- ✓ 80 object categories
- ✓ 5 captions per image
- ✓ Keypoints on 100,000 people

Collaborators

Tsung-Yi Lin Cornell Tech

Genevieve Patterson MSR

Matteo Ruggero Ronchi Caltech

Yin Cui Cornell Tech

Michael Maire TTI Chicago

Serge Belongie Cornell Tech

Lubomir Bourdev WaveOne, Inc.

Ross Girshick Facebook AI

James Hays Georgia Tech

Pietro Perona Caltech

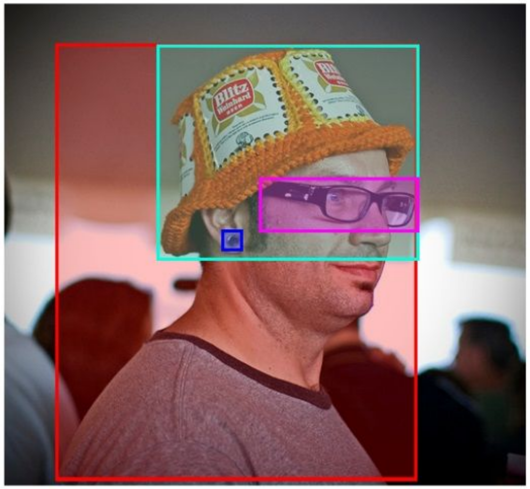
Deva Ramanan CMU

Larry Zitnick Facebook AI

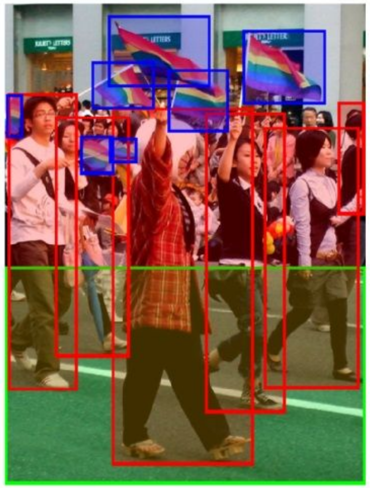
Piotr Dollár Facebook AI

Sponsors

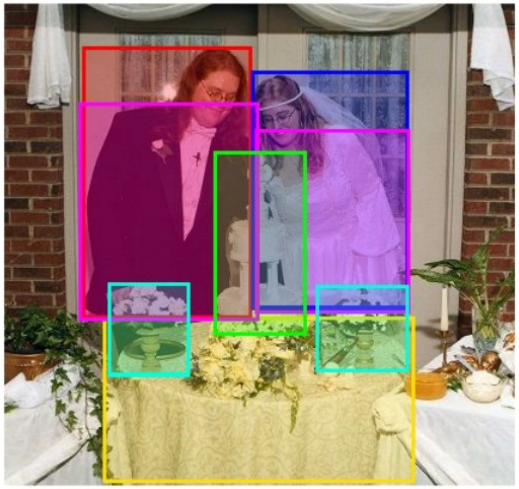
The Datasets - Flickr 30k



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

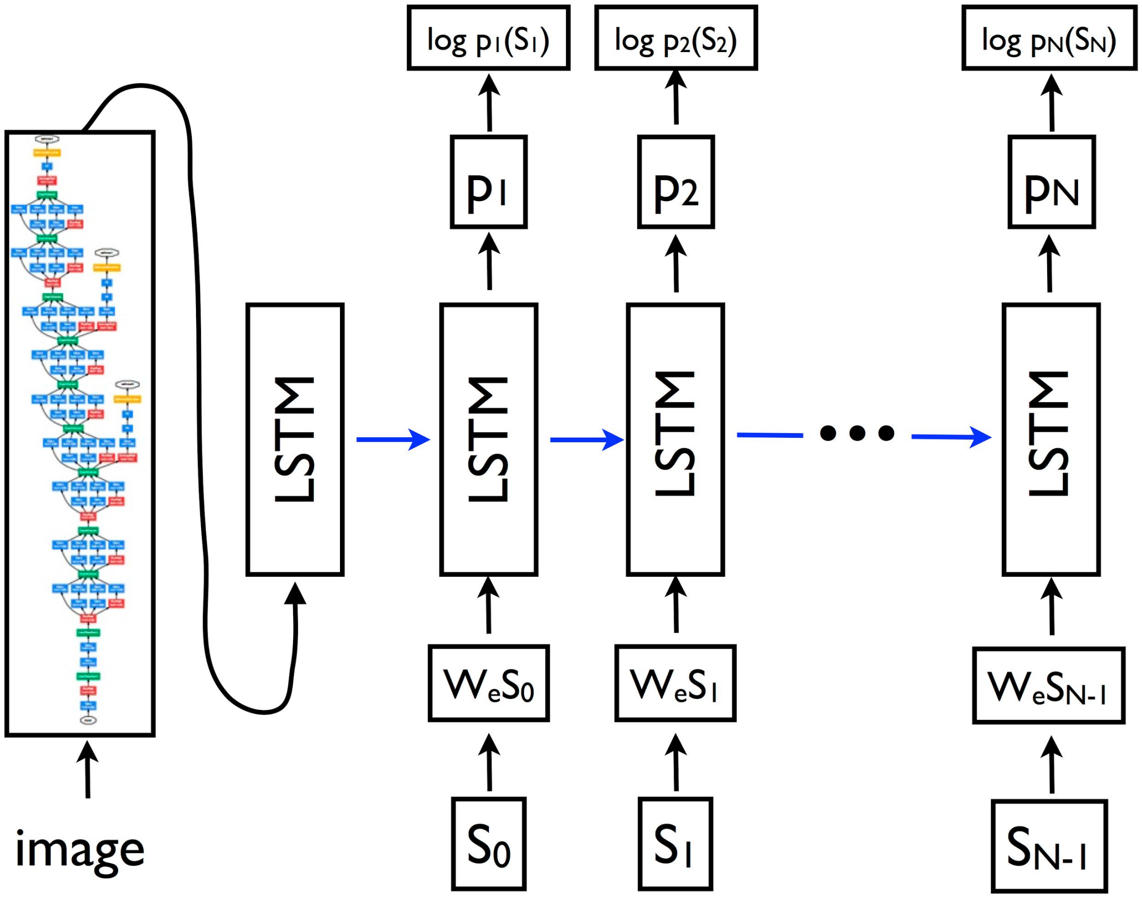


During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .



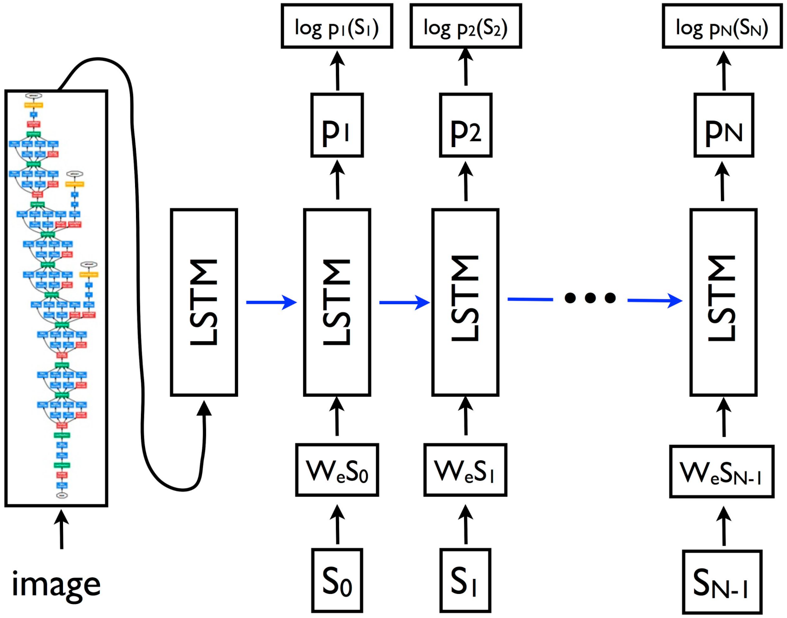
A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Simple Architecture



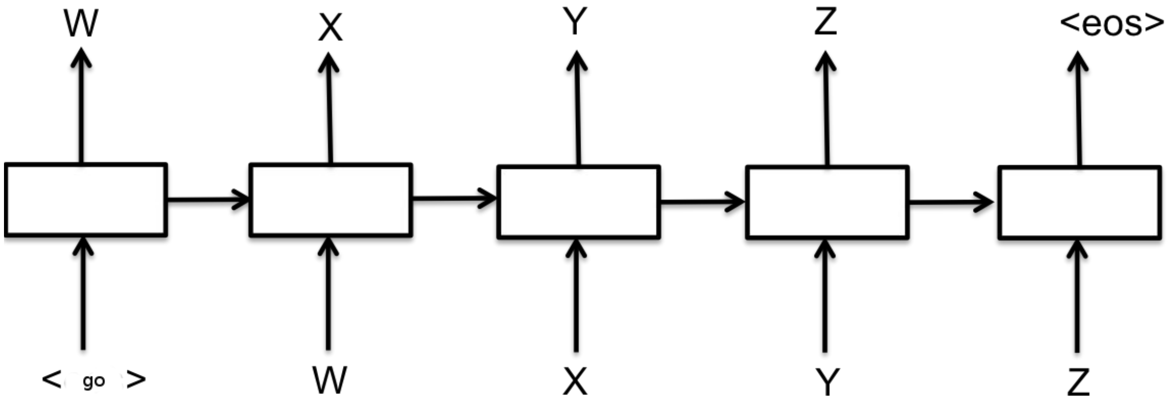
Training Vs Test time

- At train time we have the caption.
- We know what word comes next
- How do we generate a caption from just the image



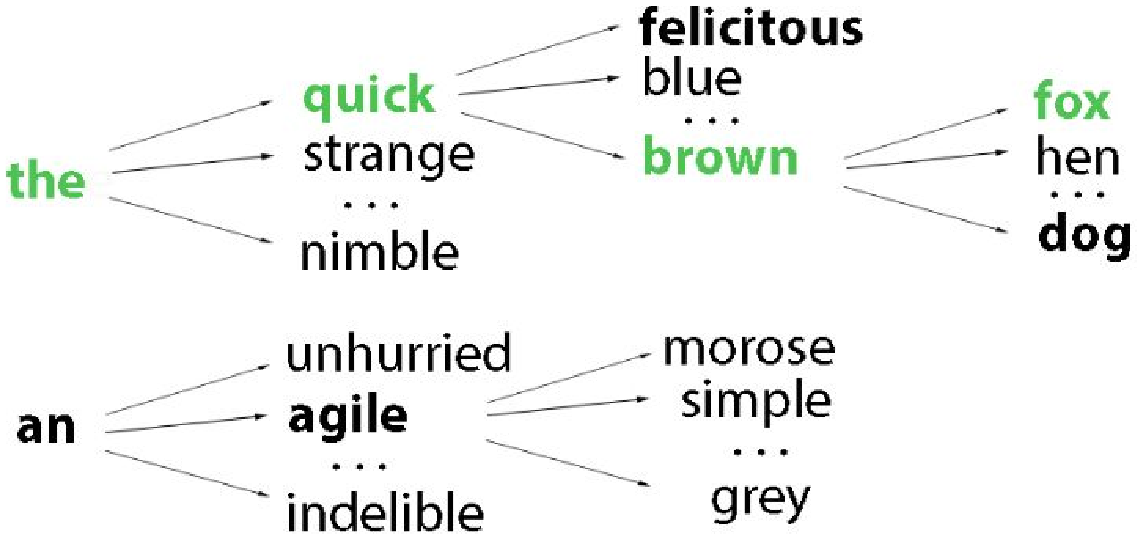
Training Vs Test time

- At test time we dynamically generate (<go> would be your image embedding)
- Predict what word comes next
- and next
- and next
- ...



Beam Search

- Instead of Greedy
- Iteratively consider k best sentences up to t
- Generate k best words for timestep t+1
- K=2 example below



Why words and pictures?

Aid the visually impaired

Why words and pictures?

Summarize Data for visual analysts

Why words and pictures

- One approach to solving AI:
    - solve each part separately
    - link up all the components afterwards
        - Stuart Russell

Why words and pictures

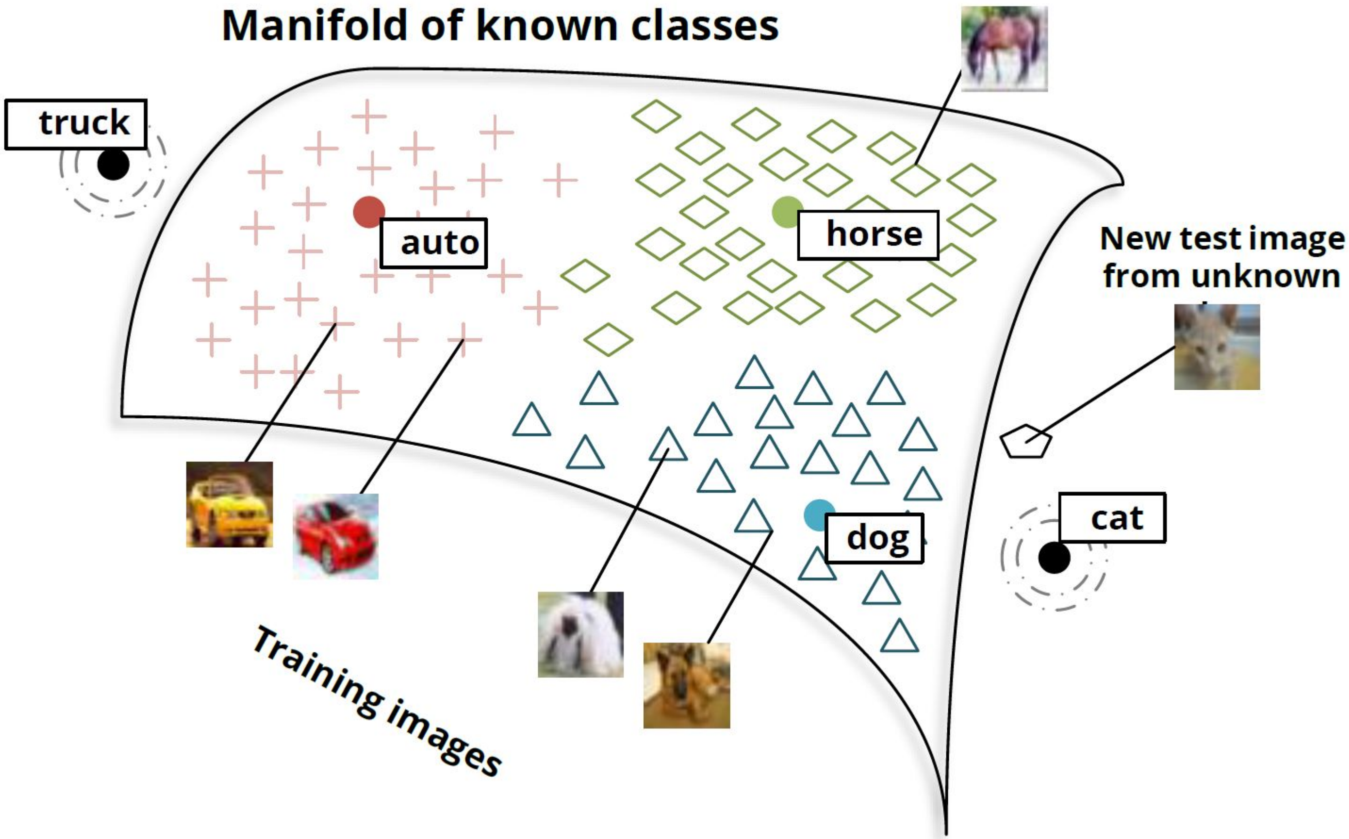
- One approach to solving AI:
  - solve each part separately
  - link up all the components afterwards
    - Stuart Russell (maybe)

Why words and pictures

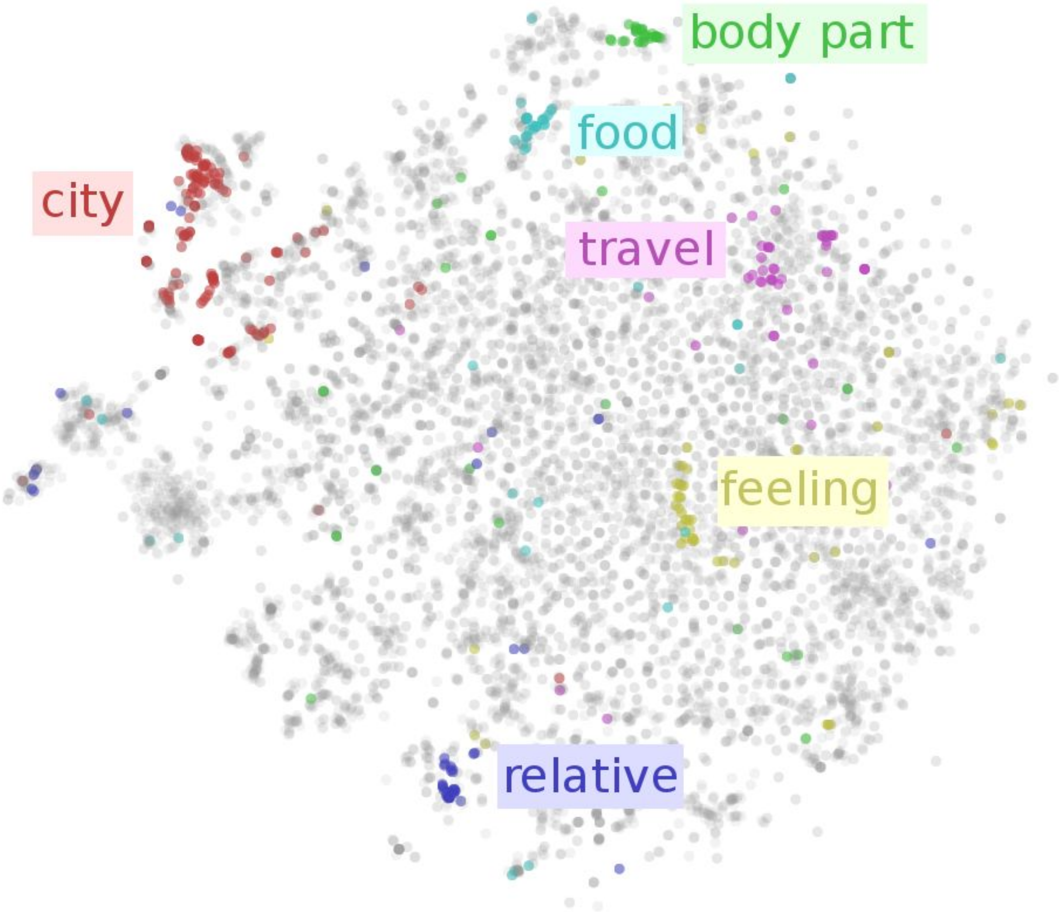
- We need some way to reconcile all the different sources of information and put them

Detour: Embedding

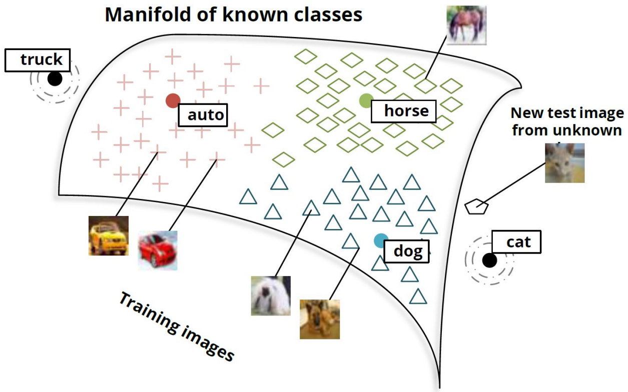
Information Representation

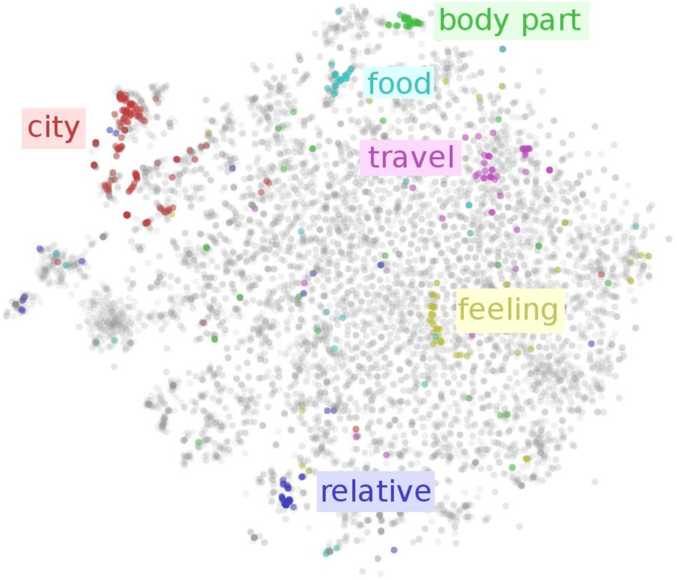


Information Representation



Information Representation





Information Representation

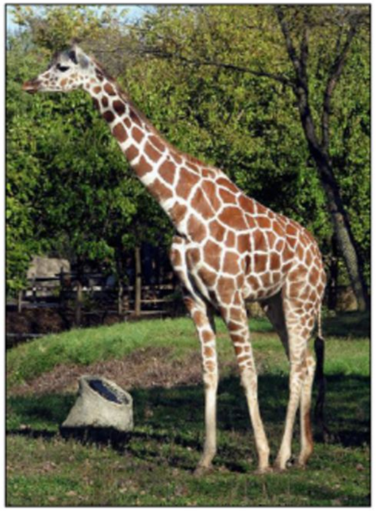
Learn an affine transformation

$$y = Wx + b$$

Caption Generation

Next steps

- Captions aren’t unique
- Captures the big picture/key points in an image



A giraffe standing next to a tree.

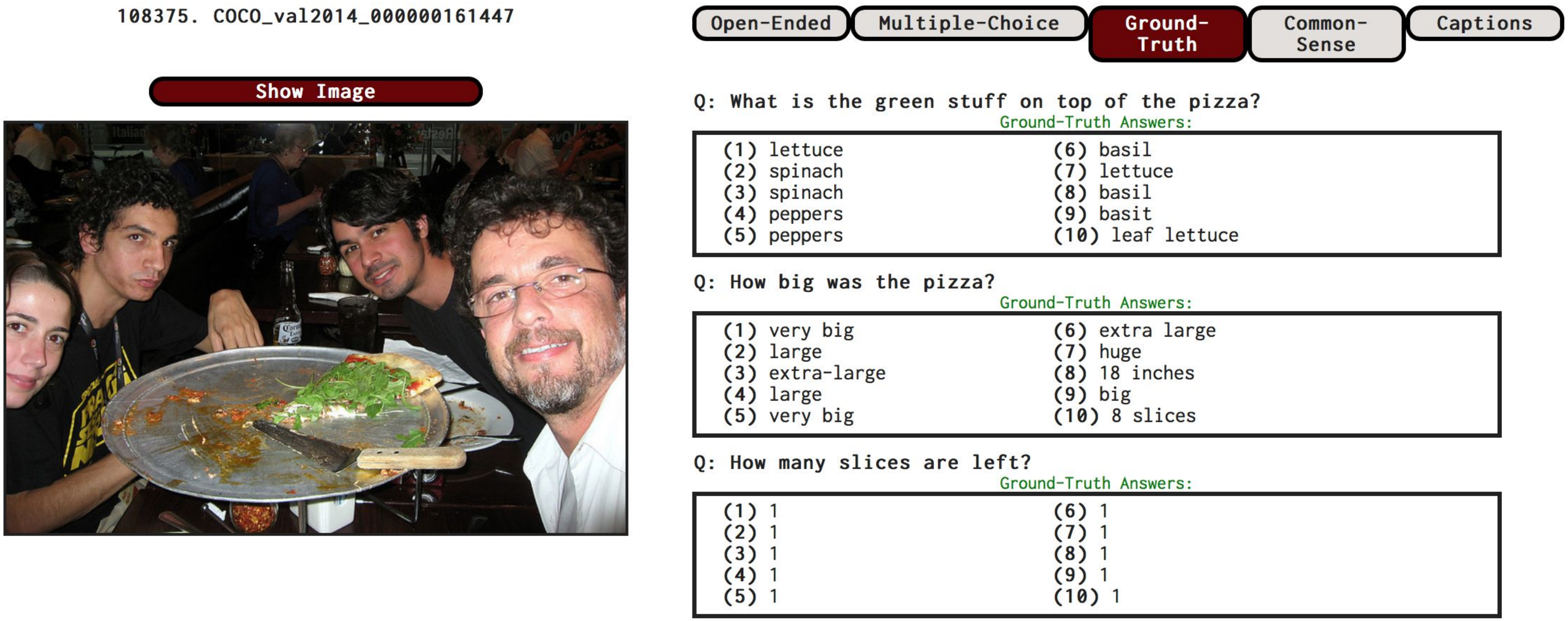
Visual Question Answering

Demo

vqa.daylen.com

The VQA 1.0 Dataset

- 200K images, 600K questions (3 per image)
- 6 million answers (10 per question)



108375. COCO_val2014_000000161447

Q: What is the green stuff on top of the pizza?

Ground-Truth Answers:

(1) lettuce
(2) spinach
(3) spinach
(4) peppers
(5) peppers
(6) basil
(7) lettuce
(8) basil
(9) basit
(10) leaf lettuce

Q: How big was the pizza?

Ground-Truth Answers:

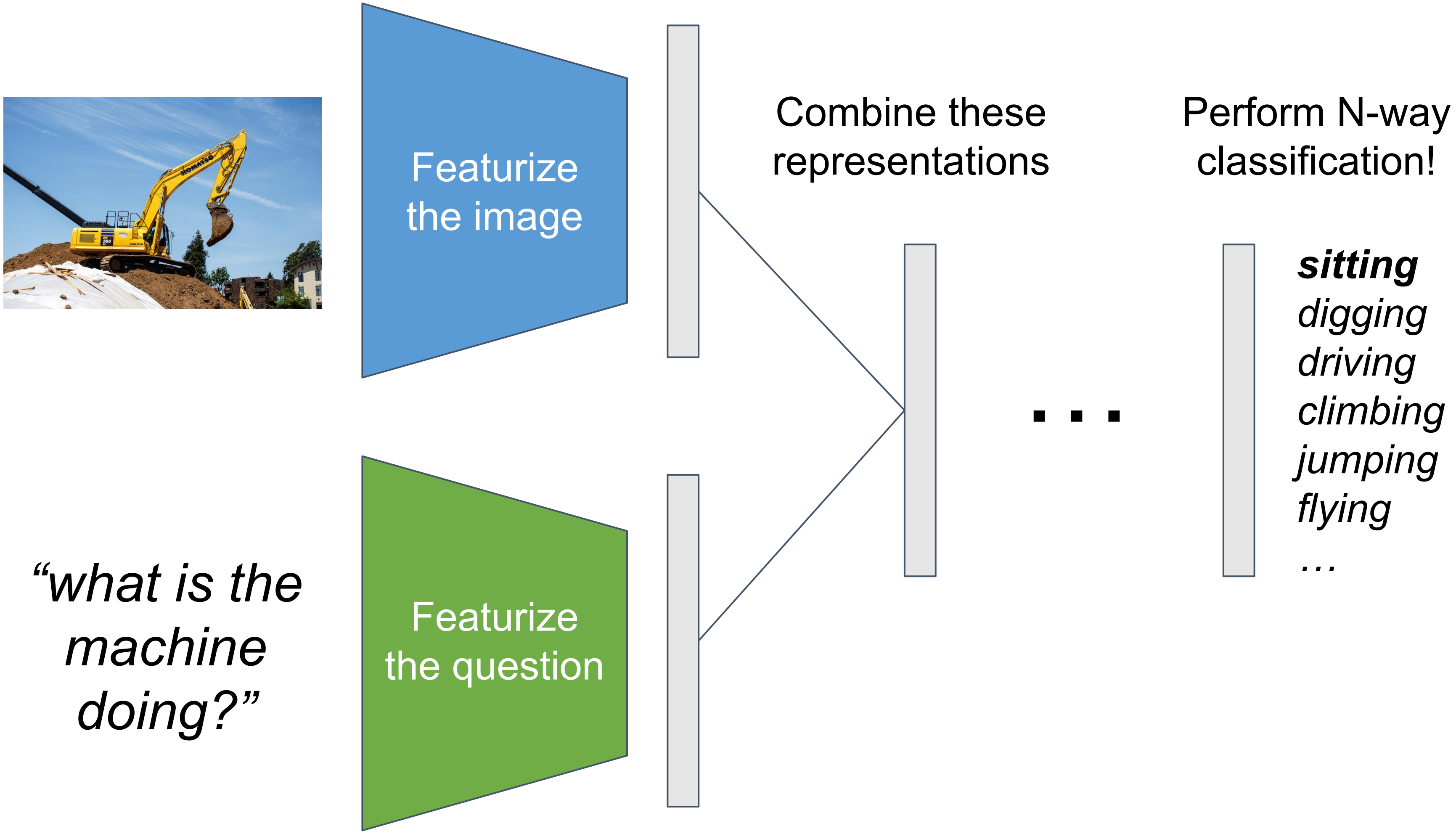
(1) very big
(2) large
(3) extra-large
(4) large
(5) very big
(6) extra large
(7) huge
(8) 18 inches
(9) big
(10) 8 slices

Q: How many slices are left?

Ground-Truth Answers:

(1) 1
(2) 1
(3) 1
(4) 1
(5) 1
(6) 1
(7) 1
(8) 1
(9) 1
(10) 1

High-level overview



Featurize the image

"what is the machine doing?"

Featurize the question

Combine these representations

. . .

Perform N-way classification!

sitting
digging
driving
climbing
jumping
flying
…

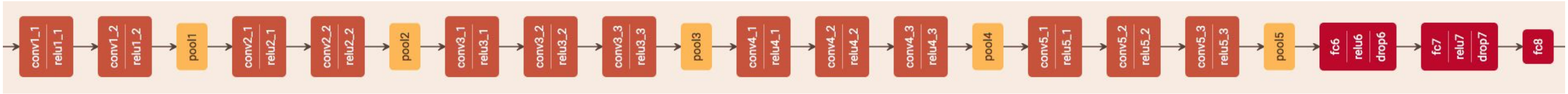
Processing the image

line, edge detectors

low level concepts

high level concepts

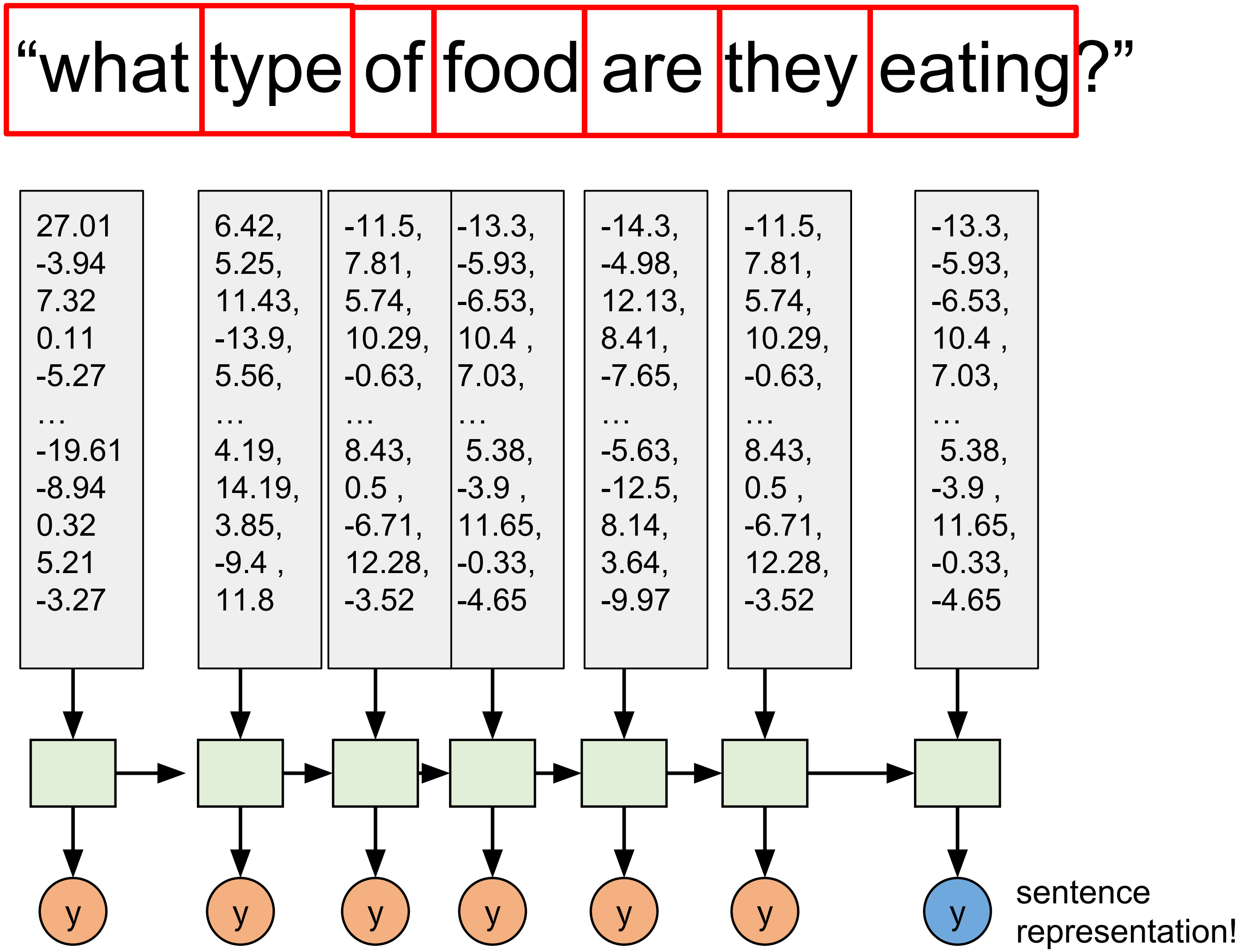
class probabilities



Solution: run the image through a state-of-the-art CNN (e.g. ResNet) and take the second-to-last layer output

Result: we have a 2048-dim vector that represents the salient aspects of the image

Processing the question



“what type of food are they eating?”

27.01	6.42,	-11.5,	-13.3,	-14.3,	-11.5,	-13.3,
-3.94	5.25,	7.81,	-5.93,	-4.98,	7.81,	-5.93,
7.32	11.43,	5.74,	-6.53,	12.13,	5.74,	-6.53,
0.11	-13.9,	10.29,	10.4 ,	8.41,	10.29,	10.4 ,
-5.27	5.56,	-0.63,	7.03,	-7.65,	-0.63,	7.03,
…	…	…	…	…	…	…
-19.61	4.19,	8.43,	5.38,	-5.63,	8.43,	5.38,
-8.94	14.19,	0.5 ,	-3.9 ,	-12.5,	0.5 ,	-3.9 ,
0.32	3.85,	-6.71,	11.65,	8.14,	-6.71,	11.65,
5.21	-9.4 ,	12.28,	-0.33,	3.64,	12.28,	-0.33,
-3.27	11.8	-3.52	-4.65	-9.97	-3.52	-4.65

sentence representation!

Now what?

Combine the image and question representations.

- Concatenation
- Eltwise Addition
- Eltwise Product
- Outer Product

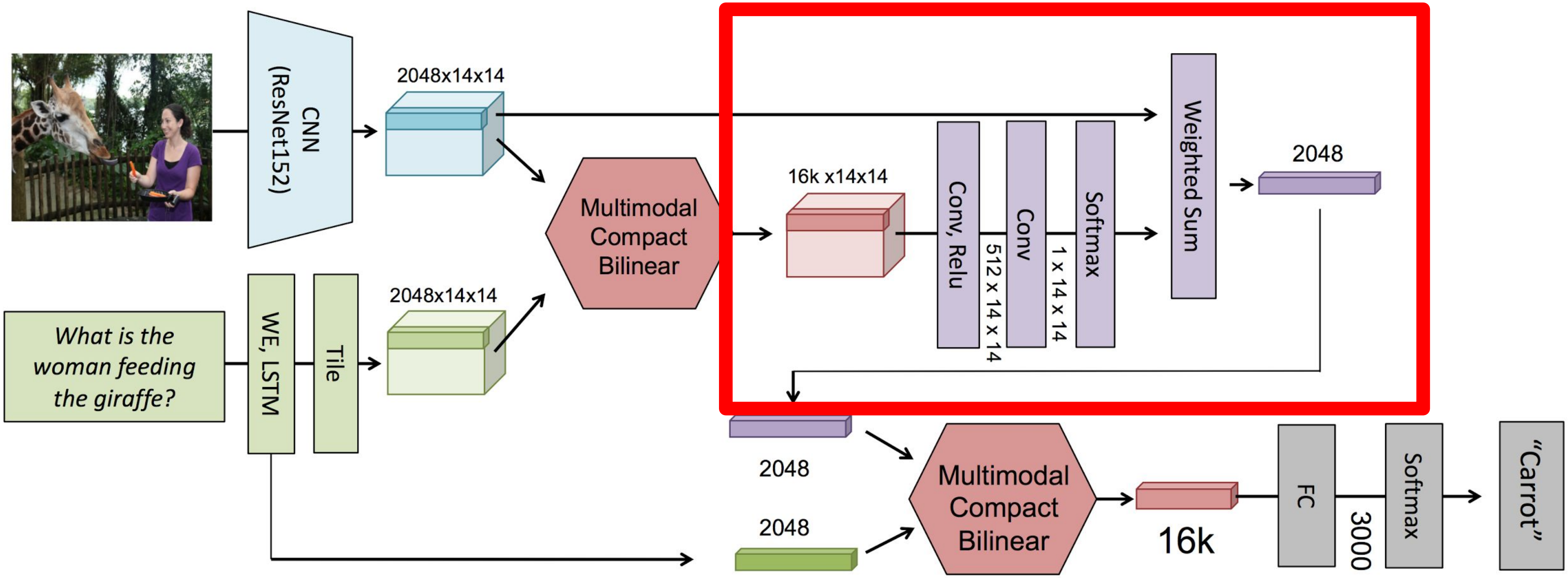
Finishing up...

Treat the problem as an N-way classification task. (We used N=3000)

Attention

Can we get a neural network to “focus on” the most important parts of the image, the way a human glances around the image when answering a question?

Attention module



Questions?