CS 2770: Computer Vision

Vision, Language, Reasoning

Prof. Adriana Kovashka
University of Pittsburgh
March 25, 2021

Plan for this lecture

- Learning the relation between images and text
  - Recurrent neural networks
  - Applications: Captioning
  - Transformers
- Reasoning: Visual question answering
  - Neuro-symbolic VQA
  - Graph convolutional networks
- Multimodal self-supervised learning

Motivation: Descriptive Text for Images

"It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin–that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns"

Scarlett O'Hara described in Gone with the Wind

Some pre-RNN good results

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

This is a picture of two dogs. The first dog is near the second furry dog.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

Kulkarni et al., CVPR 2011

Some pre-RNN bad results

Missed detections:

Here we see one potted plant.

This is a picture of one dog.

False detections:

There are one road and one cat. The furry road is in the furry cat.

This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the red road.

Incorrect attributes:

This is a photograph of two sheeps and one grass. The first black sheep is by the green grass, and by the second black sheep. The second black sheep is by the green grass.

This is a photograph of two horses and one grass. The first feathered horse is within the green grass, and by the second feathered horse. The second feathered horse is within the green grass.

Kulkarni et al., CVPR 2011

Results with Recurrent Neural Networks

"man in black shirt is playing guitar."

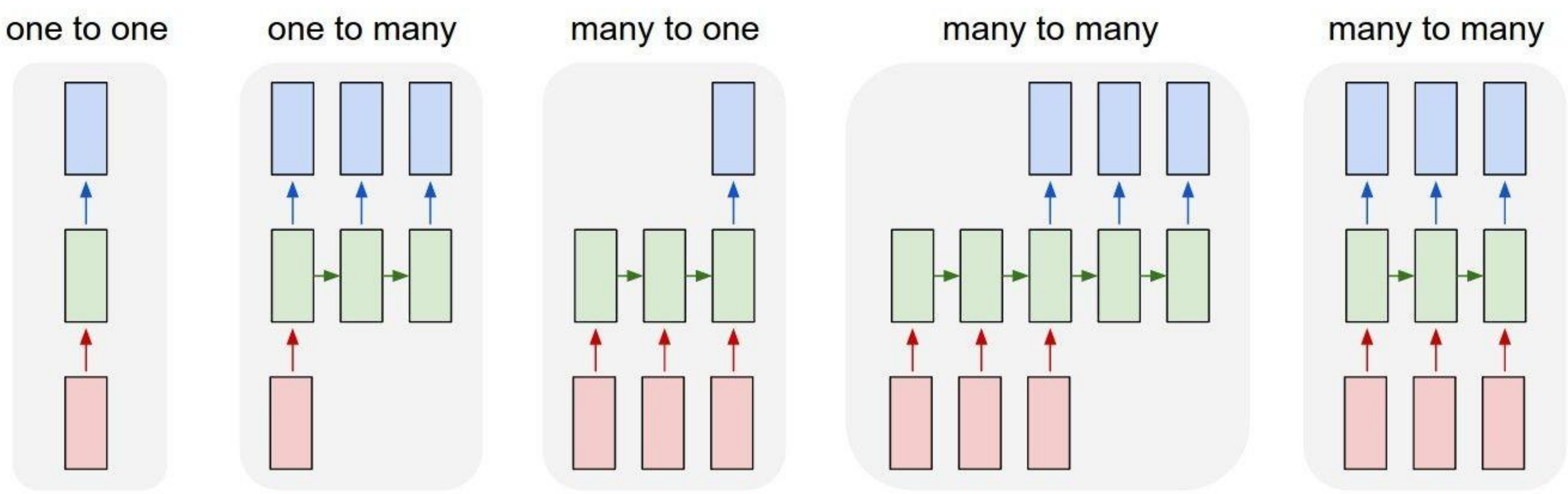
"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

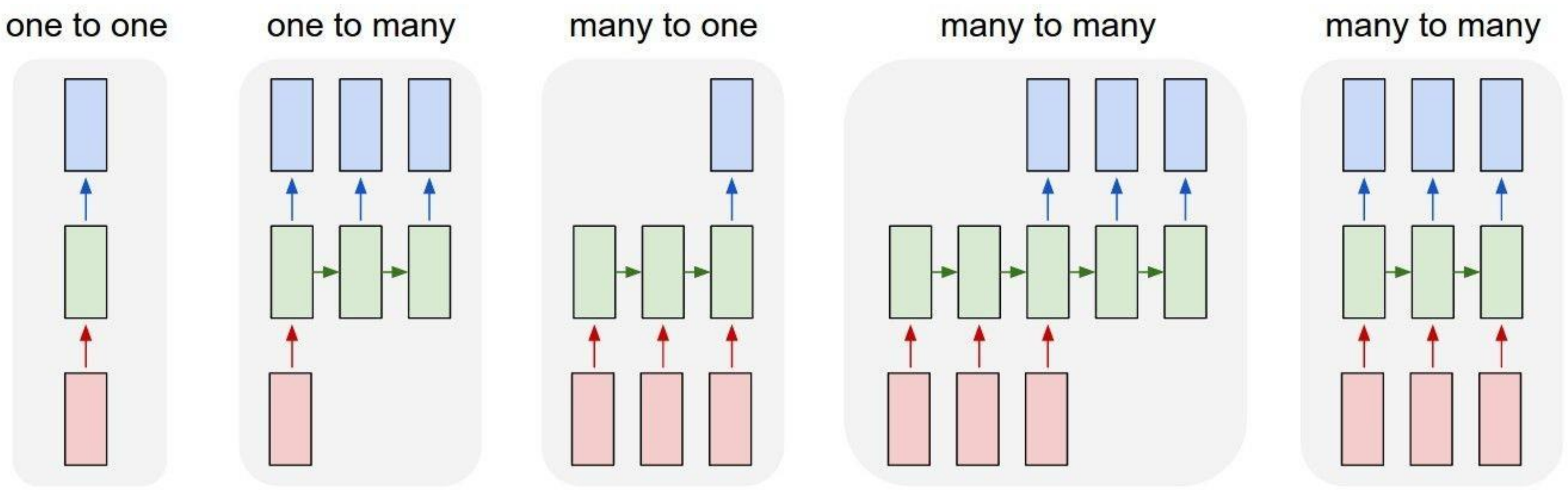
Karpathy and Fei-Fei, CVPR 2015

Recurrent Networks offer a lot of flexibility:



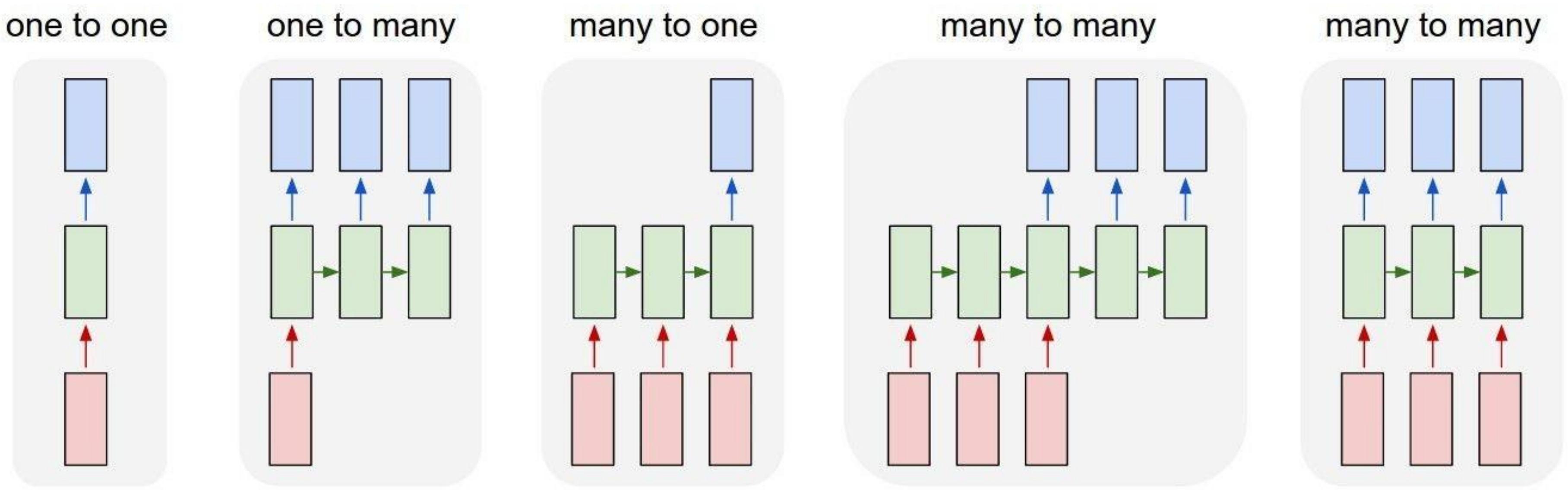
vanilla neural networks

Recurrent Networks offer a lot of flexibility:



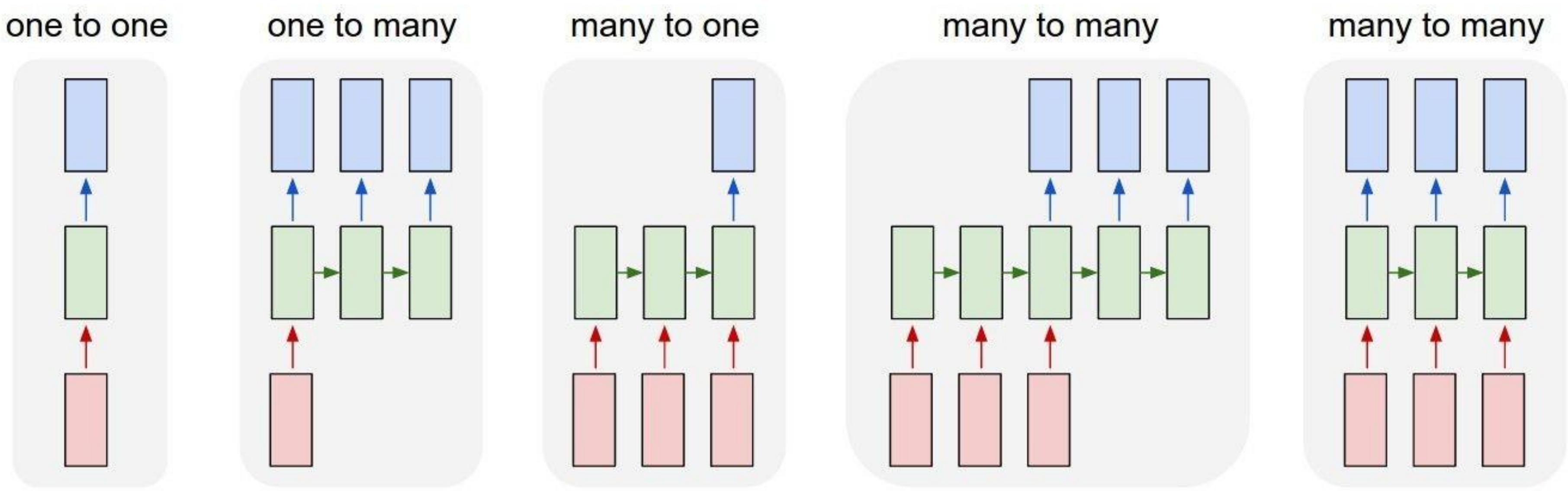
e.g. **image captioning**
image -> sequence of words

Recurrent Networks offer a lot of flexibility:



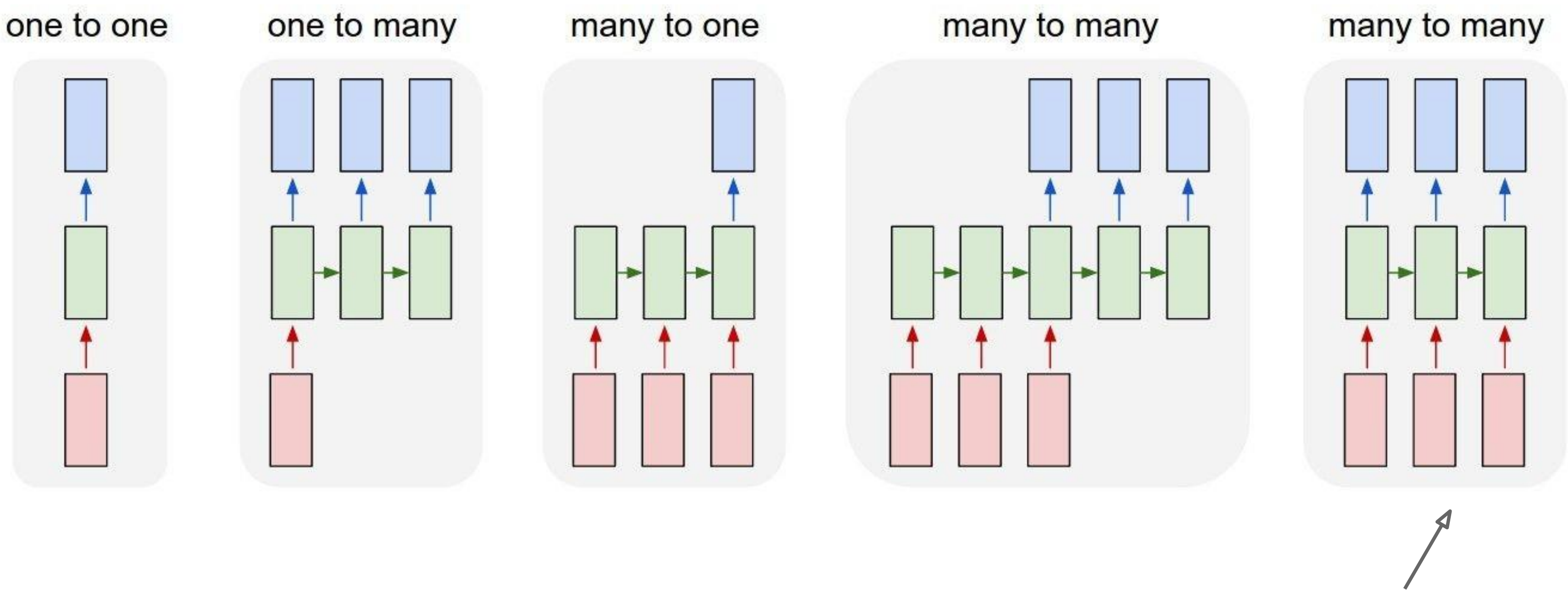
e.g. **sentiment classification**
sequence of words -> sentiment

Recurrent Networks offer a lot of flexibility:



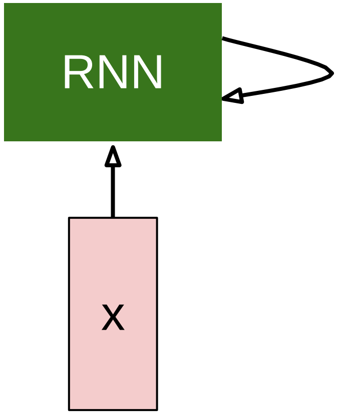
e.g. **machine translation**
seq of words -> seq of words

Recurrent Networks offer a lot of flexibility:



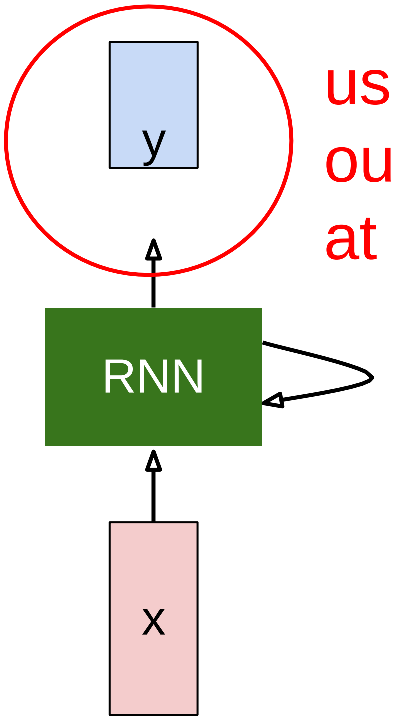
e.g. **video classification on frame level**

Recurrent Neural Network



Andrej Karpathy

Recurrent Neural Network



usually want to
output a prediction
at some time steps

Adapted from Andrej Karpathy

Recurrent Neural Network

We can process a sequence of vectors **x** by applying a recurrence formula at every time step:

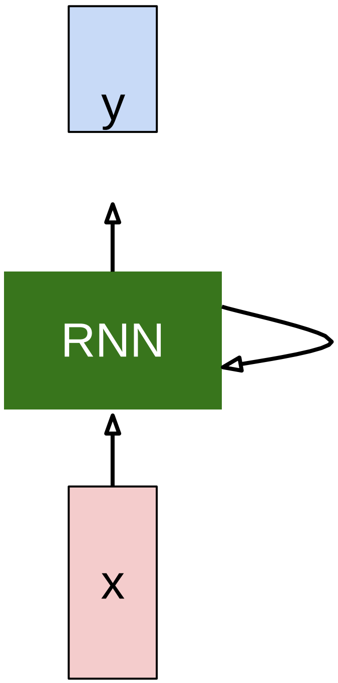
$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state

old state

input vector at some time step

some function with parameters W

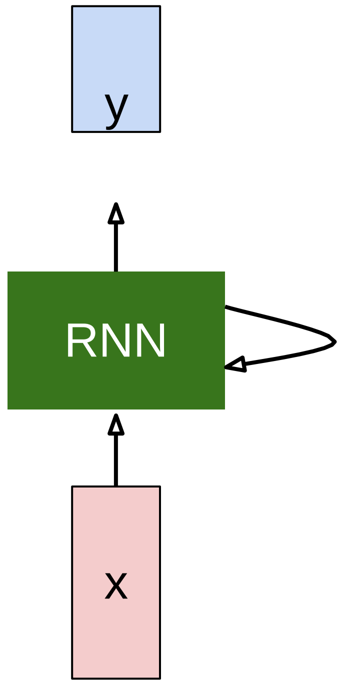


Andrej Karpathy

Recurrent Neural Network

We can process a sequence of vectors **x** by applying a recurrence formula at every time step:

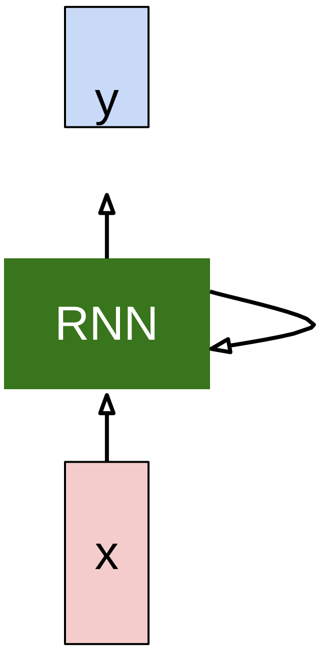
$$h_t = f_W(h_{t-1}, x_t)$$



Notice: the same function and the same set of parameters are used at every time step.

(Vanilla) Recurrent Neural Network

The state consists of a single *"hidden"* vector **h**:



$$h_t = f_W(h_{t-1}, x_t)$$

$$\downarrow$$

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

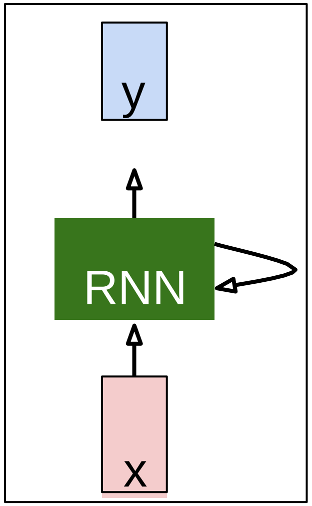
$$y_t = W_{hy}h_t$$

Example

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”

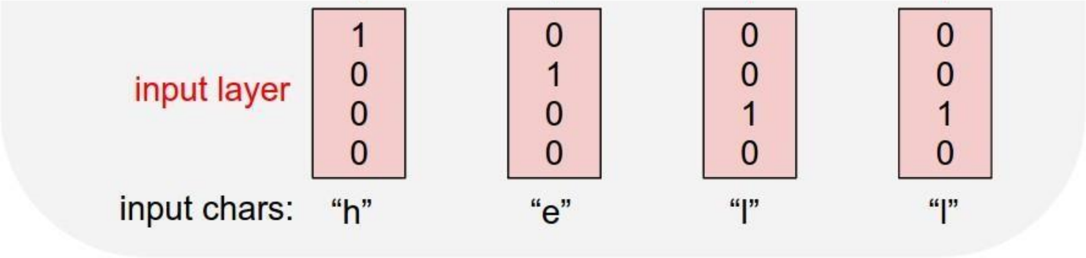


Example

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



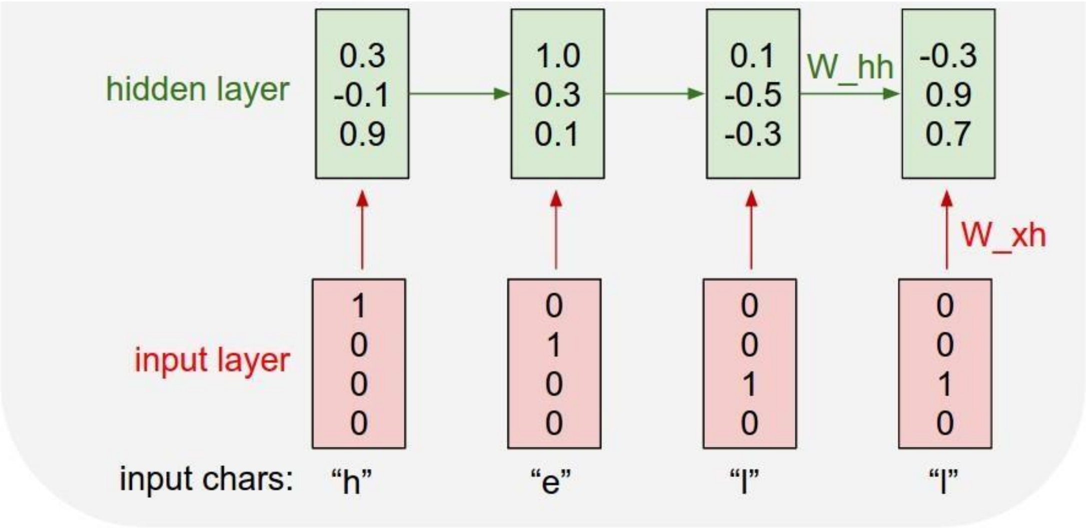
Example

Character-level language model example

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

Vocabulary: [h,e,l,o]

Example training sequence: **“hello”**



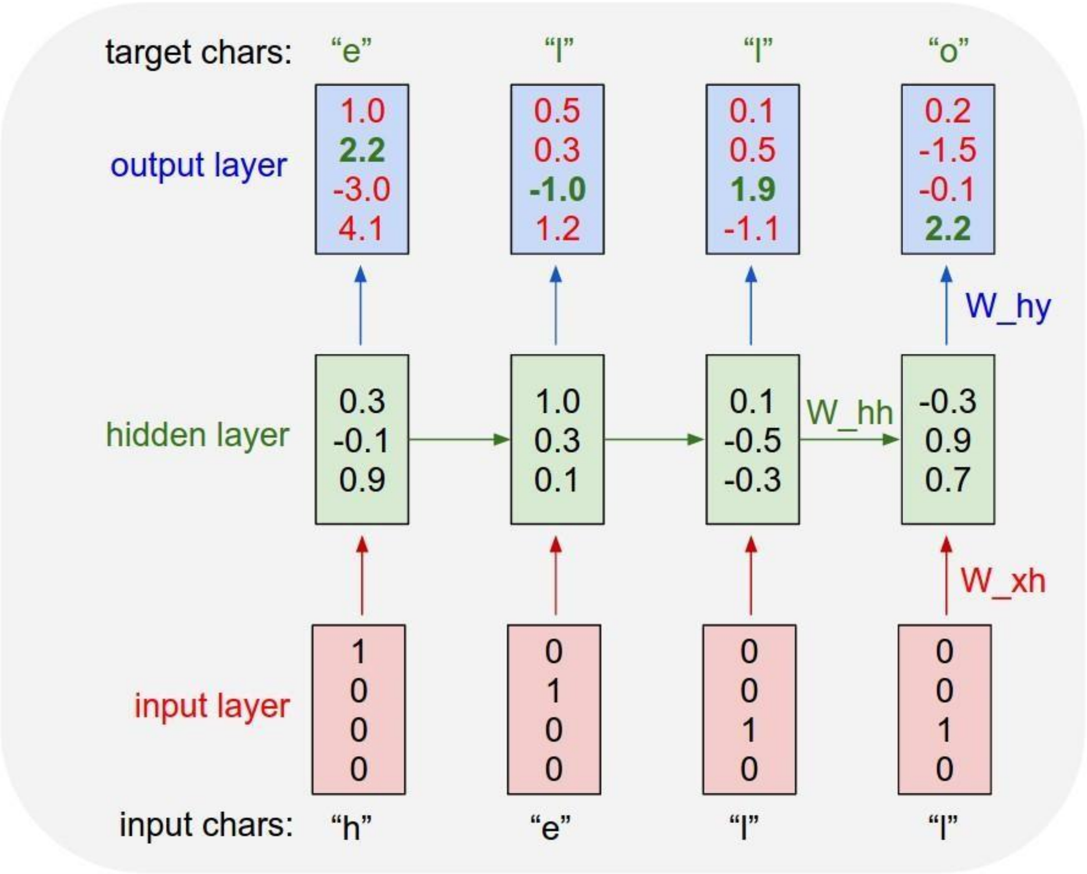
Andrej Karpathy

Example

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
"hello"

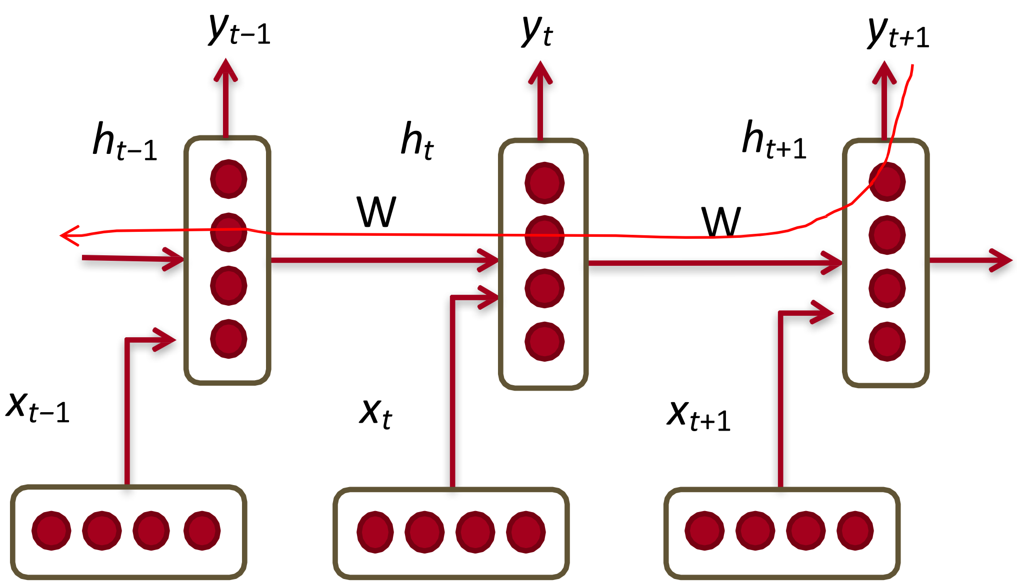


Loss?

Cross-entropy for every time step (generate token that really comes next)

The vanishing gradient problem

- The error at a time step ideally can tell a previous time step from many steps away to change during backprop
- But we’re multiplying together many values between 0 and 1



Adapted from Richard Socher

The vanishing gradient problem

- Total error is the sum of each error at time steps t

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$

- Chain rule:

$$\frac{\partial E_t}{\partial W} = \sum_{k=1}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

- More chain rule:

$$\frac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}}$$

- Derivative of vector wrt vector is a Jacobian matrix of partial derivatives; norm of this matrix can become very small or very large quickly [Bengio et al 1994, Pascanu et al 2013], leading to vanishing/exploding gradient

Adapted from Richard Socher

The vanishing gradient problem for language models

- In the case of language modeling or question answering words from time steps far away are not taken into consideration when training to predict the next word

- Example:

  Jane walked into the room. John walked in too. It was late in the day. Jane said hi to _____

Gated Recurrent Units (GRUs)

- More complex hidden unit computation in recurrence!
- Introduced by Cho et al. 2014
- Main ideas:
  - keep around memories to capture long distance dependencies
  - allow error messages to flow at different strengths depending on the inputs

Gated Recurrent Units (GRUs)

- Standard RNN computes hidden layer at next time step directly: $$h_t = f\left(W^{(hh)}h_{t-1} + W^{(hx)}x_t\right)$$
- GRU first computes an update **gate** (another layer) based on current input word vector and hidden state

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right)$$

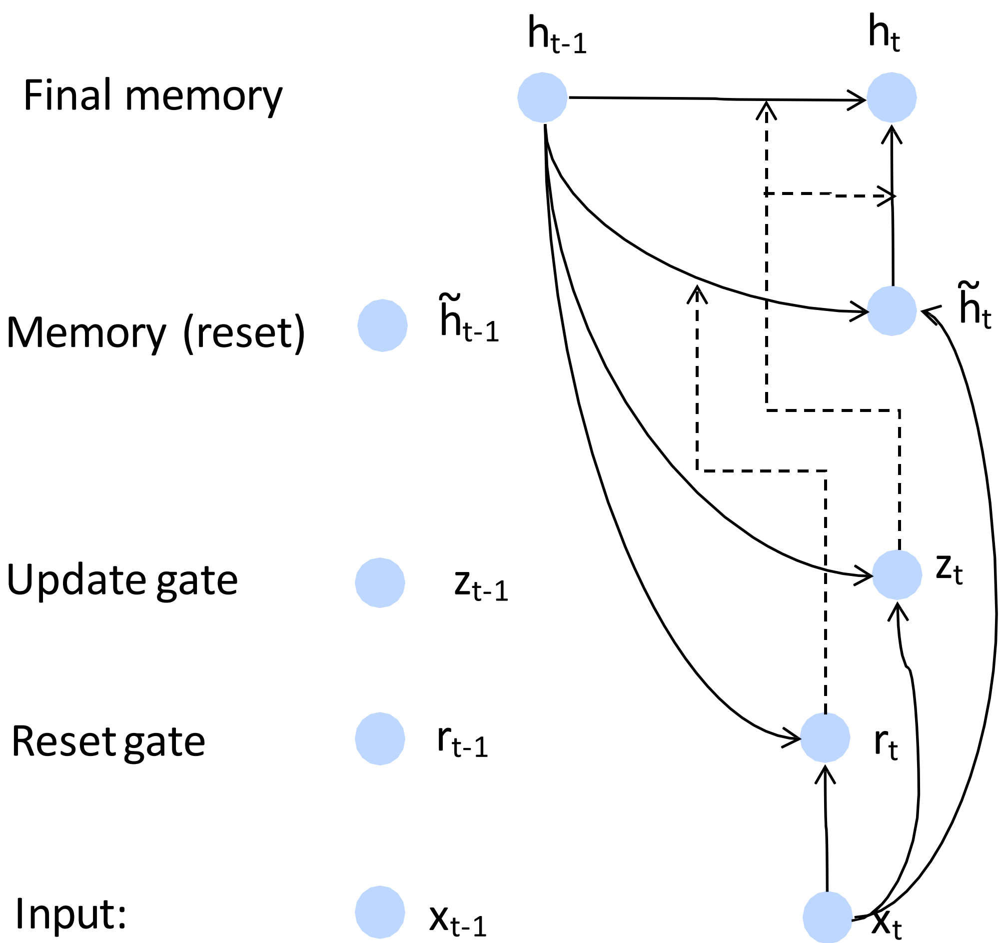
- Compute reset gate similarly but with different weights

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right)$$

Gated Recurrent Units (GRUs)

- Update gate $z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right)$
- Reset gate $r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right)$
- New memory content: $\tilde{h}_t = \tanh\left(Wx_t + r_t \circ Uh_{t-1}\right)$
  If reset gate unit is ~0, then this ignores previous memory and only stores the new word information
- Final memory at time step combines current and previous time steps: $h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t$

Gated Recurrent Units (GRUs)



$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right)$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right)$$

$$\tilde{h}_t = \tanh\left(Wx_t + r_t \circ Uh_{t-1}\right)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t$$

x_t

h_{t-1}

r_t

z_t

Gated Recurrent Units (GRUs)

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right)$$
$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right)$$
$$\tilde{h}_t = \tanh\left(Wx_t + r_t \circ Uh_{t-1}\right)$$
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t$$

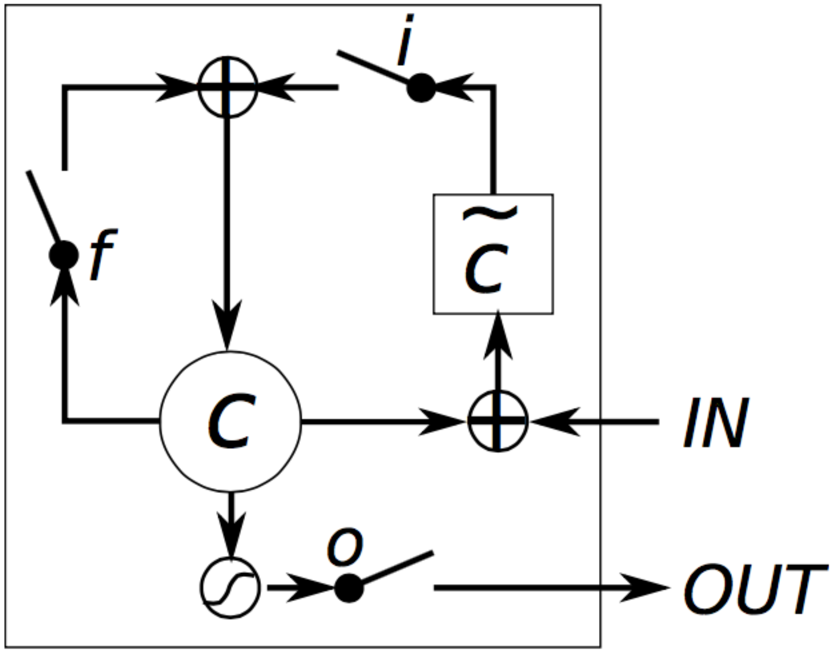
- If reset is close to 0, ignore previous hidden state
  - Allows model to drop information that is irrelevant in the future
- Update gate z controls how much of past state should matter now
  - If z close to 1, then we can copy information in that unit through many time steps!
  - **Less vanishing gradient**!

Long-short-term-memories (LSTMs)

- Proposed by Hochreiter and Schmidhuber in 1997
- We can make the units even more complex
- Allow each time step to modify
  - Input gate (current cell matters) $i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right)$
  - Forget (gate 0, forget past) $f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1}\right)$
  - Output (how much cell is exposed) $o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right)$
  - New memory cell $\tilde{c}_t = \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-1}\right)$
- Final memory cell: $c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$
- Final hidden state: $h_t = o_t \circ \tanh(c_t)$

Adapted from Richard Socher

Long-short-term-memories (LSTMs)

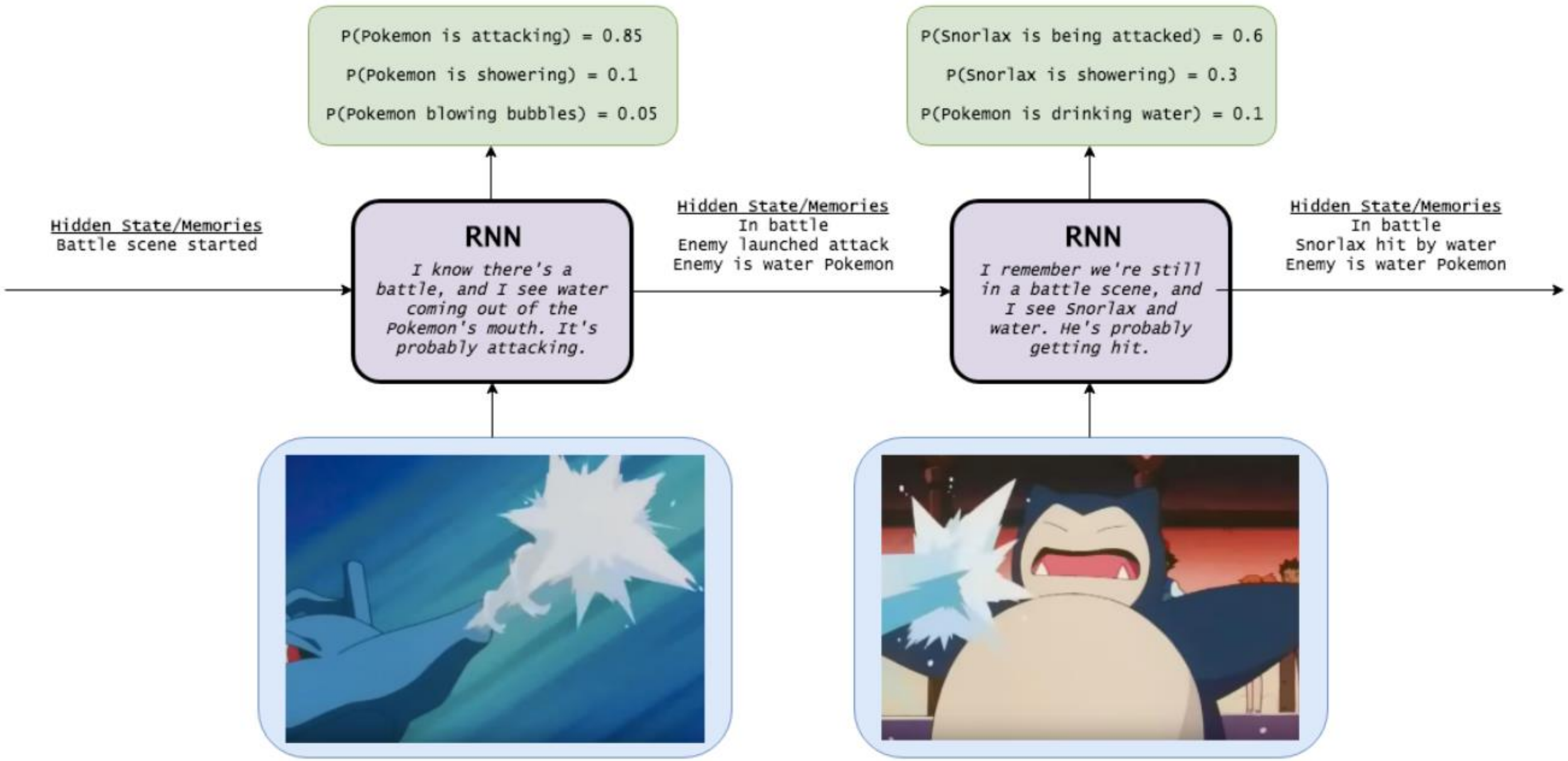


Intuition: memory cells can keep information intact, unless inputs makes them forget it or overwrite it with new input

Cell can decide to output this information or just store it

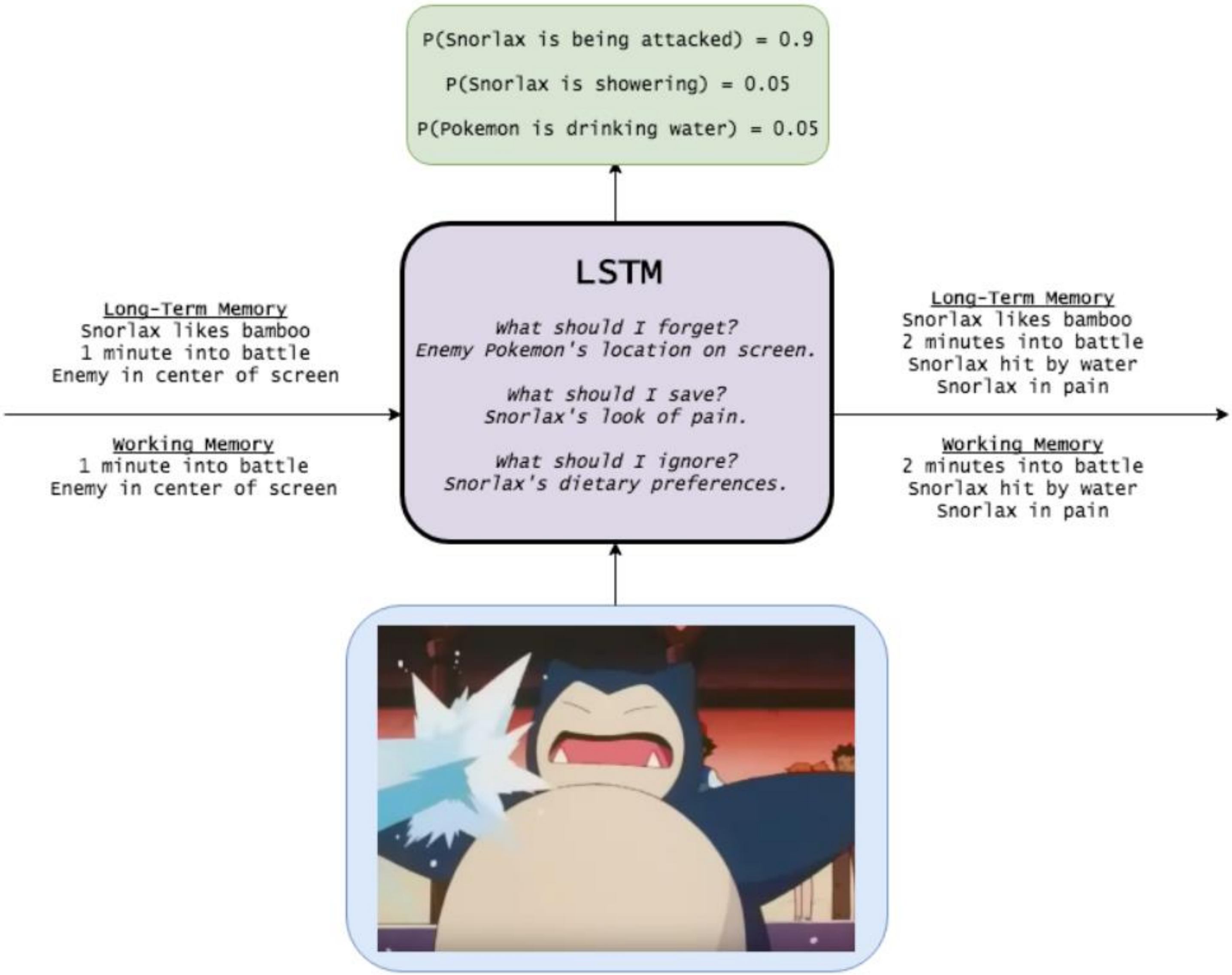
Richard Socher, figure from wildml.com

Long-short-term-memories (LSTMs)



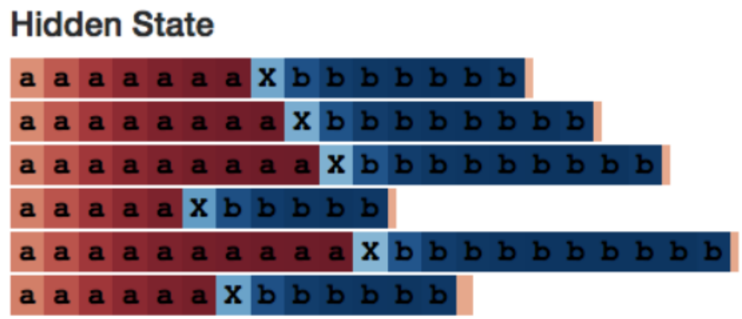
http://blog.echen.me/2017/05/30/exploring-lstms/

Long-short-term-memories (LSTMs)



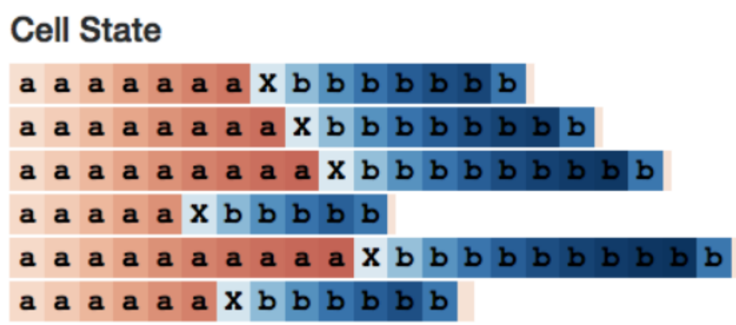
http://blog.echen.me/2017/05/30/exploring-lstms/

LSTM for counting



I built a small web app to play around with LSTMs, and Neuron #2 seems to be counting both the number of a's it's seen, as well as the number of b's. (Remember that cells are shaded according to the neuron's activation, from dark red [-1] to dark blue [+1].)

What about the cell state? It behaves similarly:



One interesting thing is that the working memory looks like a "sharpened" version of the long-term memory. Does this hold true in general?

It does. (This is exactly as we would expect, since the long-term memory gets squashed by the tanh activation function and the output gate limits what gets passed on.) For example, here is an overview of all 10 cell state nodes at once. We see plenty of light-colored cells, representing values close to 0.

http://blog.echen.me/2017/05/30/exploring-lstms/

Plan for this lecture

- Learning the relation between images and text
  - Recurrent neural networks
  - Applications: Captioning
  - Transformers
- Reasoning: Visual question answering
  - Neuro-symbolic VQA
  - Graph convolutional networks
- Multimodal self-supervised learning

Generating poetry with RNNs

Sonnet 116 – Let me not ...

by William Shakespeare

Let me not to the marriage of true minds
Admit impediments. Love is not love
Which alters when it alteration finds,
Or bends with the remover to remove:
O no! it is an ever-fixed mark
That looks on tempests and is never shaken;
It is the star to every wandering bark,
Whose worth's unknown, although his height be taken.
Love's not Time's fool, though rosy lips and cheeks
Within his bending sickle's compass come:
Love alters not with his brief hours and weeks,
But bears it out even to the edge of doom.
If this be error and upon me proved,
I never writ, nor no man ever loved.

Andrej Karpathy

Generating poetry with RNNs

at first:

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

train more

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

train more

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

train more

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

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

Andrej Karpathy

Generating poetry with RNNs

```
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.
```

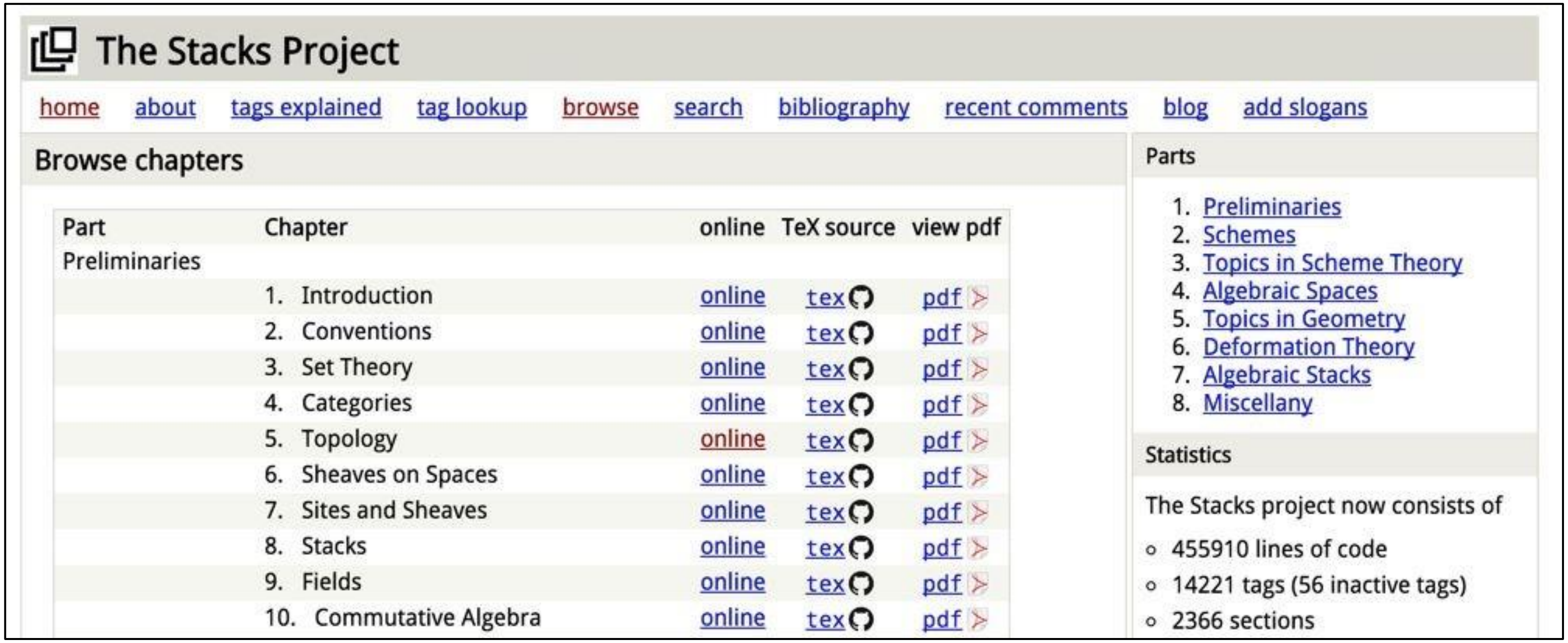
```
VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
```

Andrej Karpathy

Generating textbooks with RNNs

open source textbook on algebraic geometry



Latex source

Andrej Karpathy

Generating textbooks with RNNs

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 $X_{spaces, étale}$ 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 $U \cap U = \coprod_{i=1,\ldots,n} U_i$ 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 $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

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. □

Generating textbooks with RNNs

Proof. Omitted. $\square$

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{étale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. $\square$

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma **??**. $\square$

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

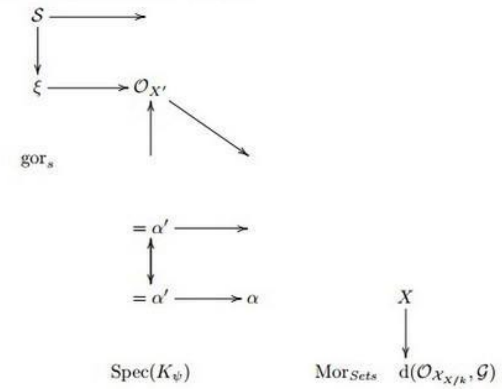
be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. $\square$

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.

$\square$

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. $\square$

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

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

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. $\square$

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.

Andrej Karpathy

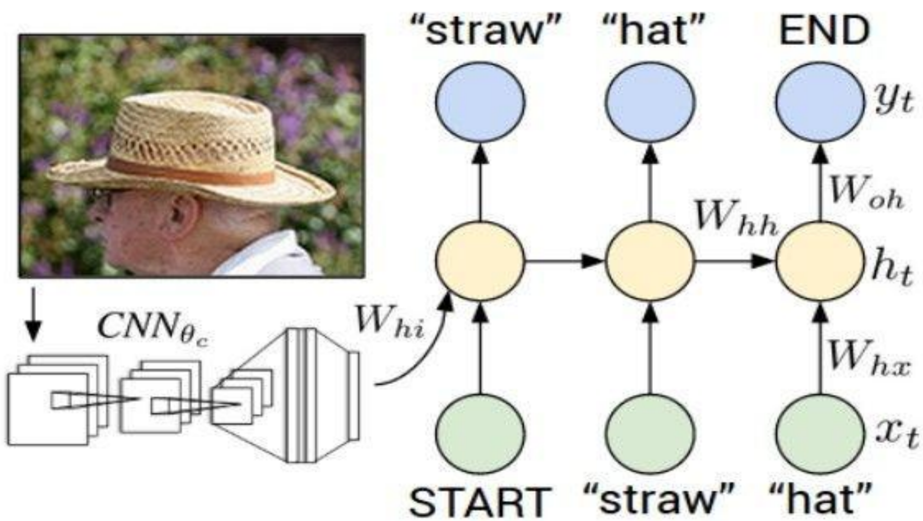
Generating code with RNNs

```c
static void do_command(struct seq_file *m, void *v)
{
  int column = 32 << (cmd[2] & 0x80);
  if (state)
    cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
  else
    seq = 1;
  for (i = 0; i < 16; i++) {
    if (k & (1 << 1))
      pipe = (in_use & UMXTHREAD_UNCCA) +
        ((count & 0x00000000fffffff8) & 0x000000f) << 8;
    if (count == 0)
      sub(pid, ppc_md.kexec_handle, 0x20000000);
    pipe_set_bytes(i, 0);
  }
  /* Free our user pages pointer to place camera if all dash */
  subsystem_info = &of_changes[PAGE_SIZE];
  rek_controls(offset, idx, &soffset);
  /* Now we want to deliberately put it to device */
  control_check_polarity(&context, val, 0);
  for (i = 0; i < COUNTER; i++)
    seq_puts(s, "policy ");
}
```

Generated C code

Andrej Karpathy

Image Captioning



CVPR 2015:
Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei
Show and Tell: A Neural Image Caption Generator, Vinyals et al.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.
Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick

Adapted from Andrej Karpathy

Image Captioning

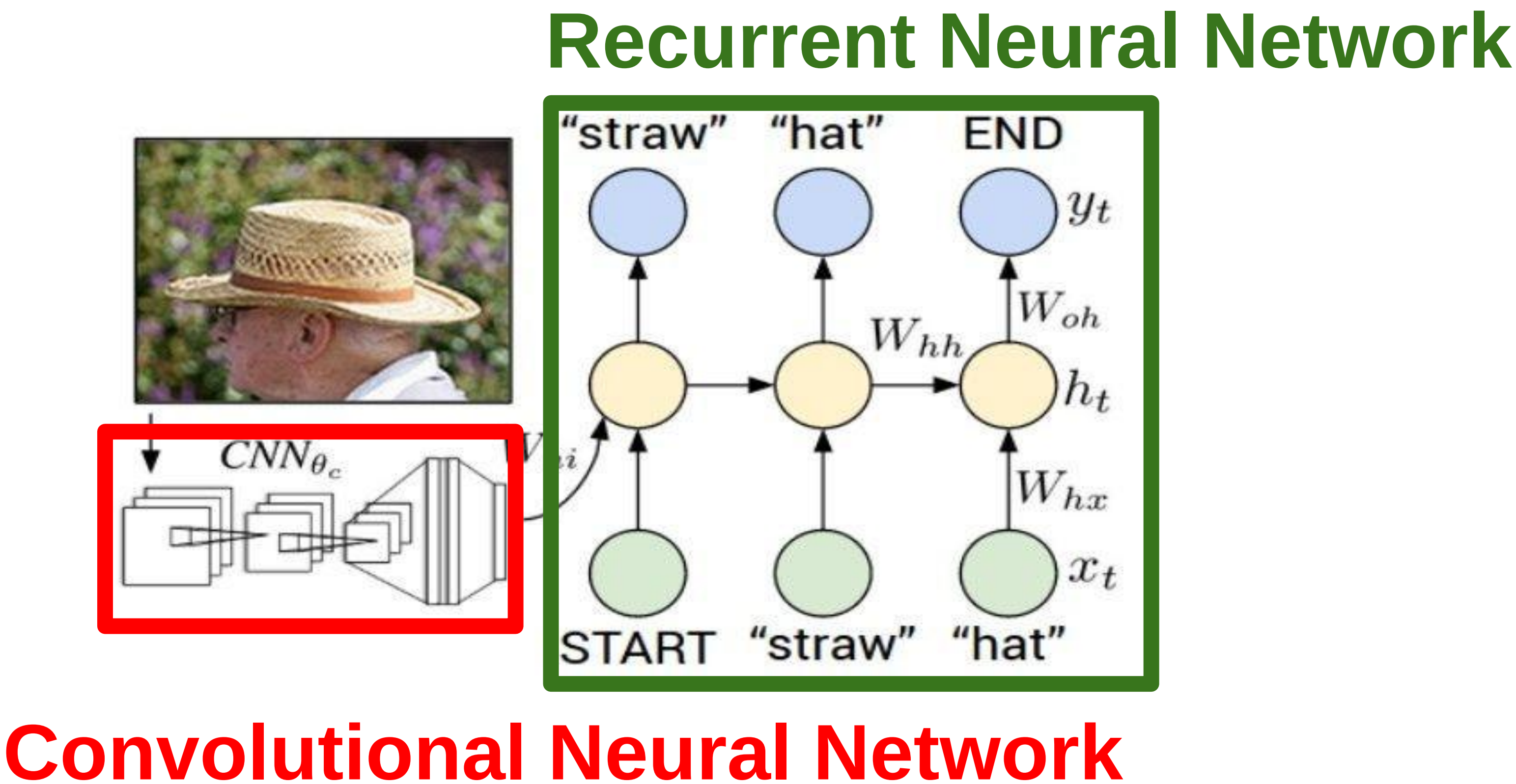
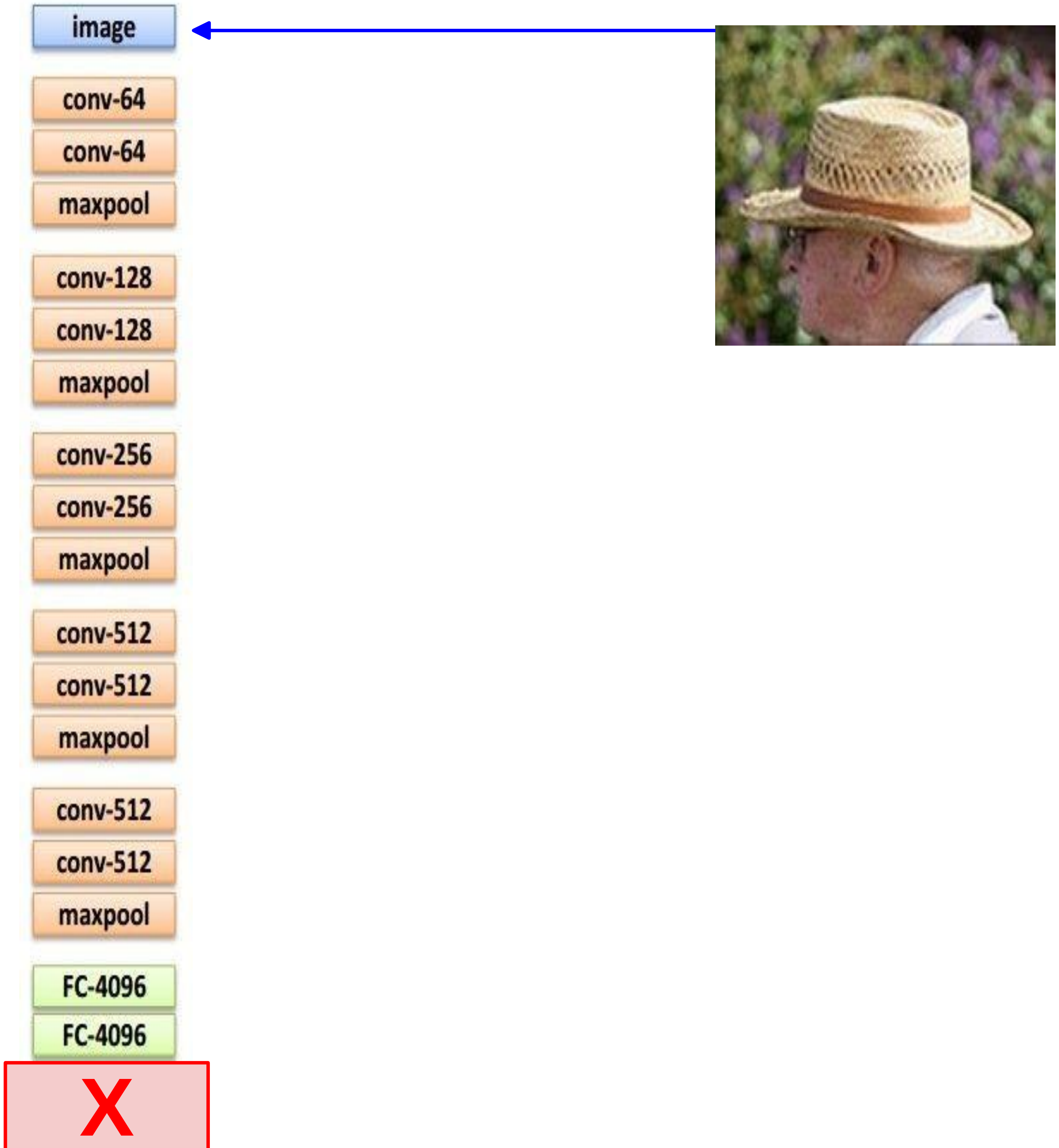


Image Captioning

test image

test image

Andrej Karpathy



test image

Andrej Karpathy

Image Captioning

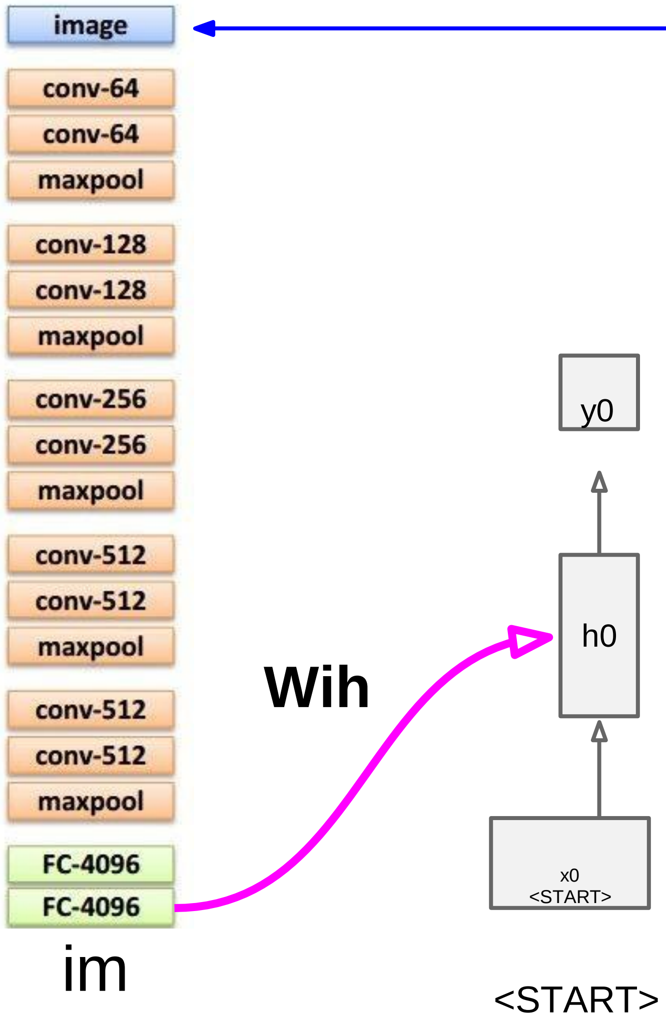
test image

x0
<START>

<START>

Andrej Karpathy

Image Captioning



test image

before:

$h = \tanh(W_{xh} * x + W_{hh} * h)$

now:

$h = \tanh(W_{xh} * x + W_{hh} * h + \mathbf{W_{ih} * im})$

Image Captioning

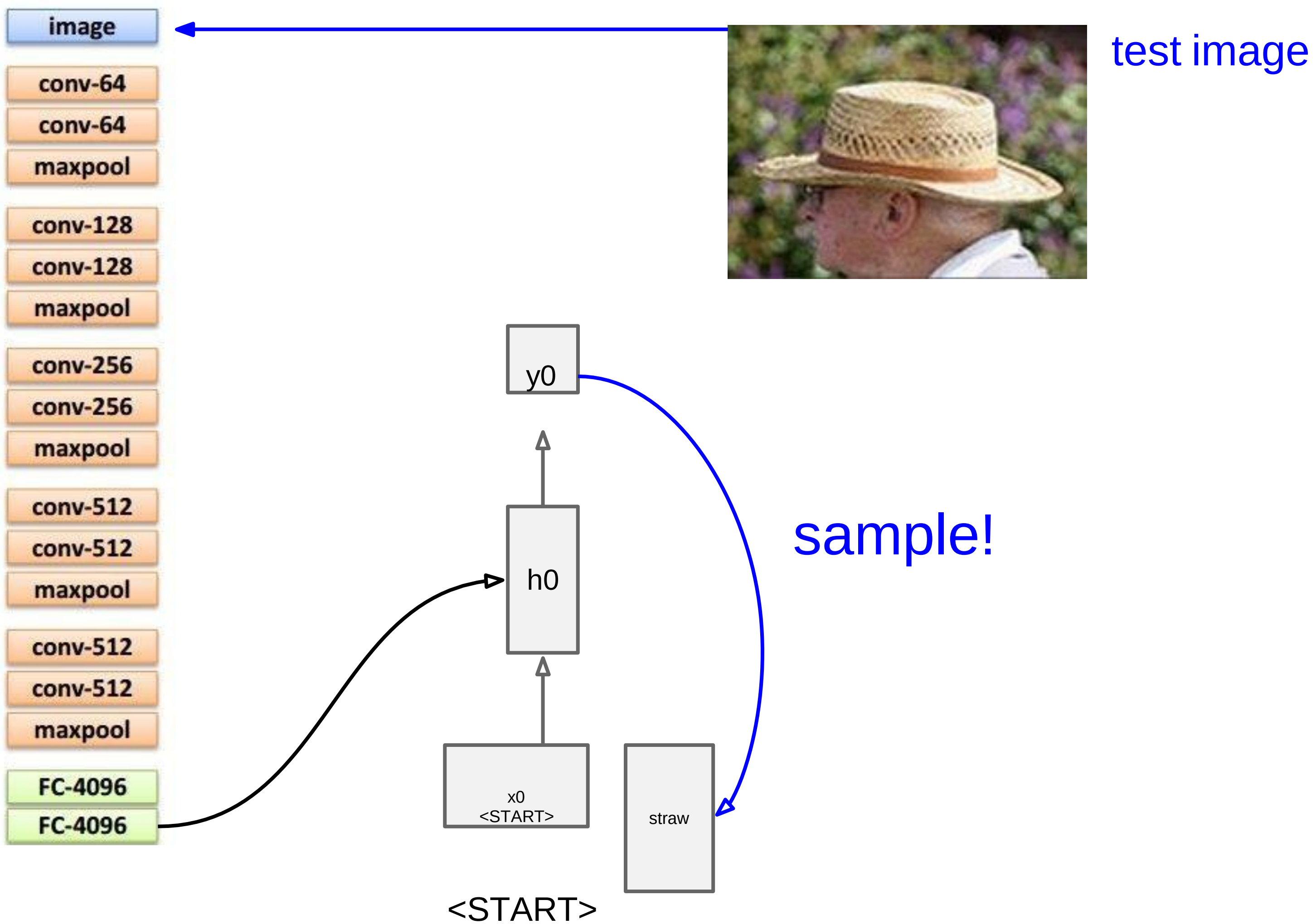


Image Captioning

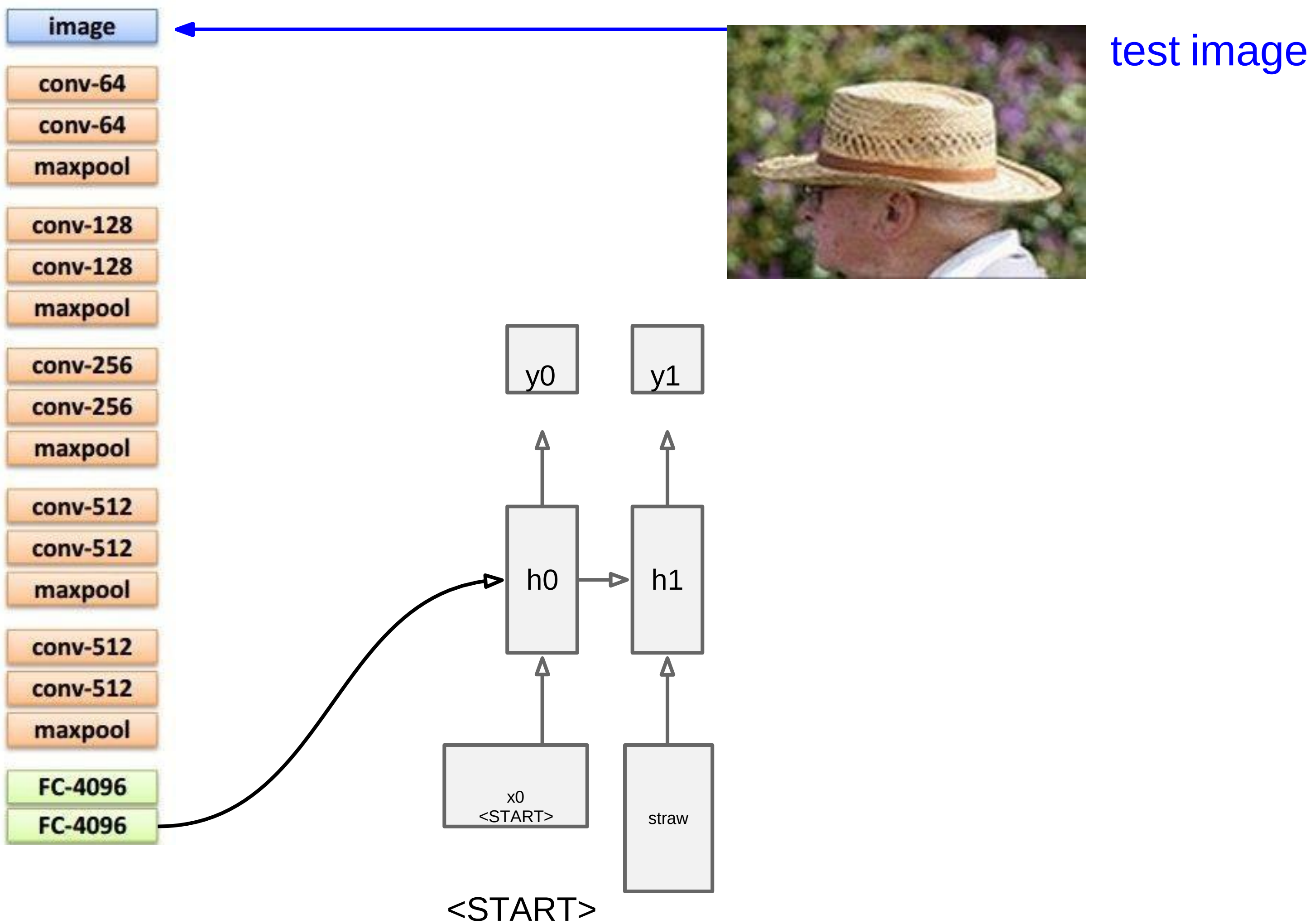
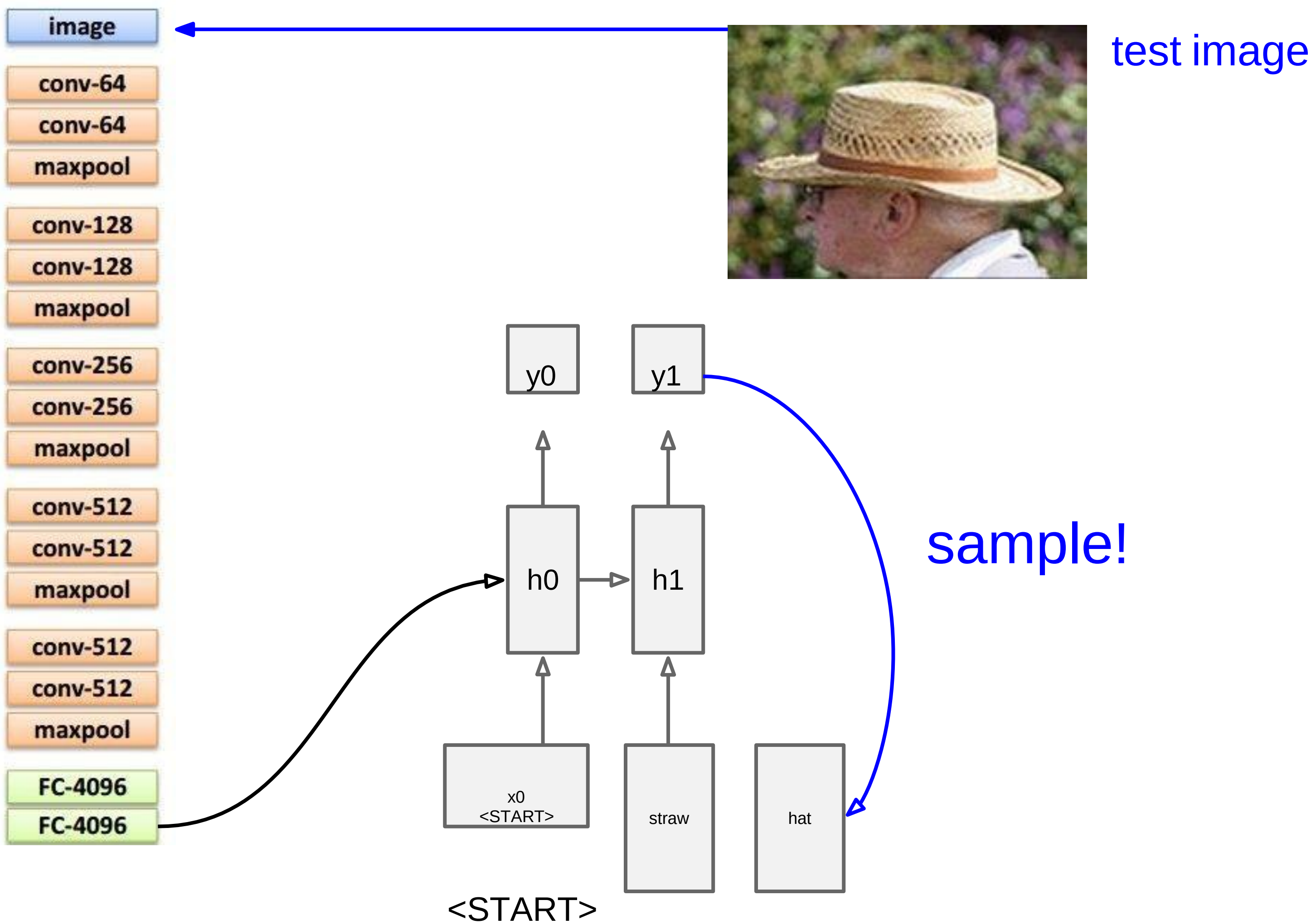


Image Captioning



Andrej Karpathy

Image Captioning

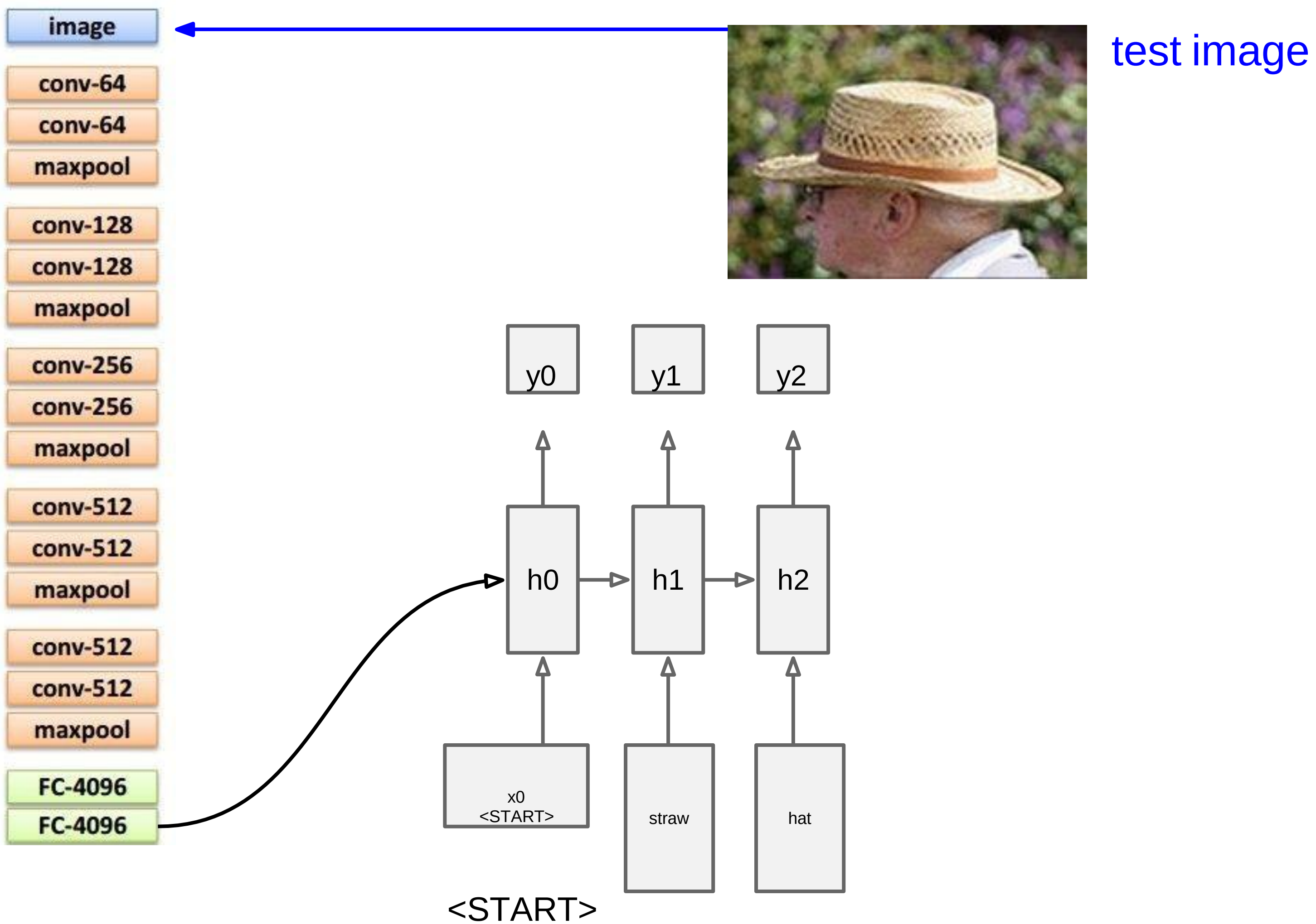
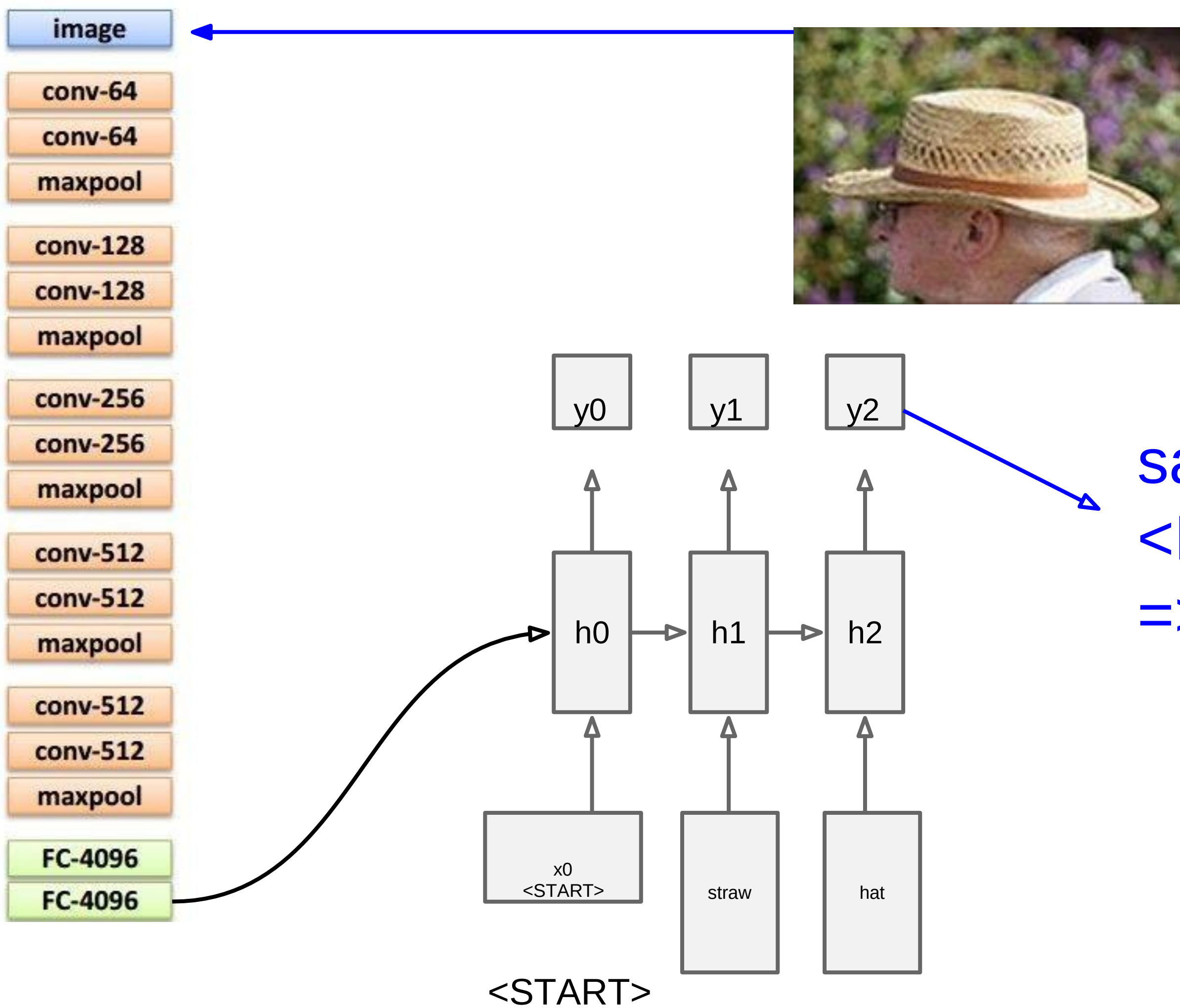


Image Captioning



test image

Caption generated:
“straw hat”

sample
<END> token
=> finish.

Adapted from Andrej Karpathy

Image Captioning

"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."

Andrej Karpathy

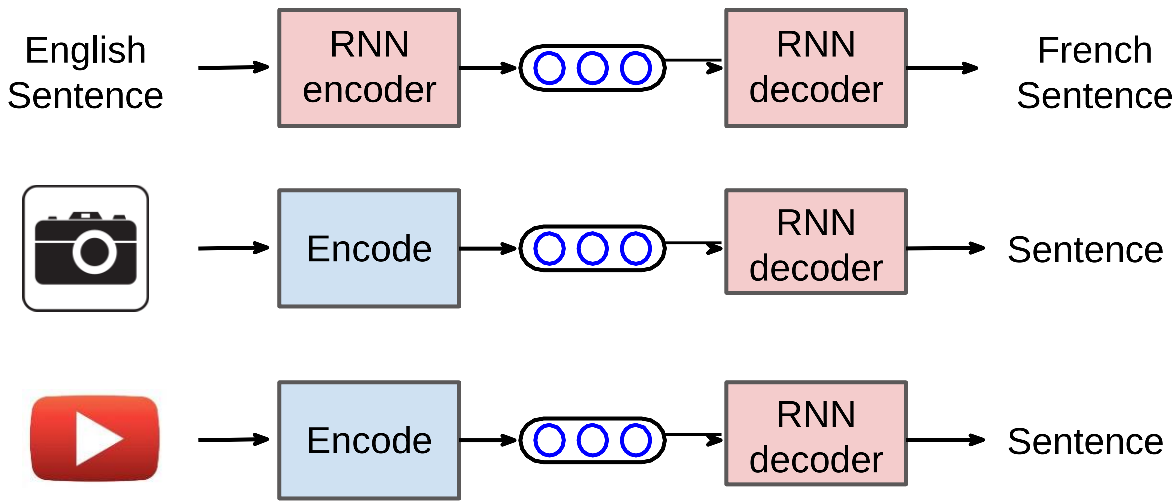
Video Captioning

Generate descriptions for events depicted in video clips

A monkey pulls a dog’s tail and is chased by the dog.

Venugopalan et al., “Translating Videos to Natural Language using Deep Recurrent Neural Networks”, NAACL-HTL 2015

Video Captioning



[Sutskever et al. NIPS'14]

[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

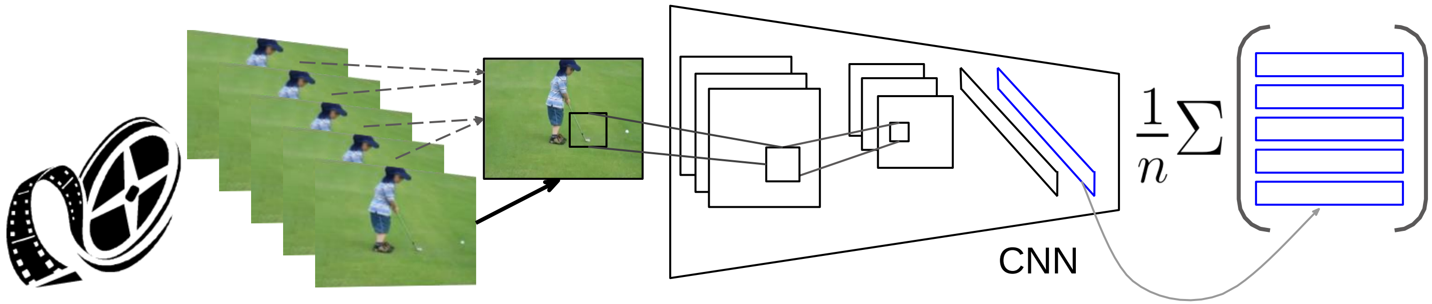
[Venugopalan et. al. NAACL'15] (this work)

Key Insight:

Generate feature representation of the video and "decode" it to a sentence

Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning



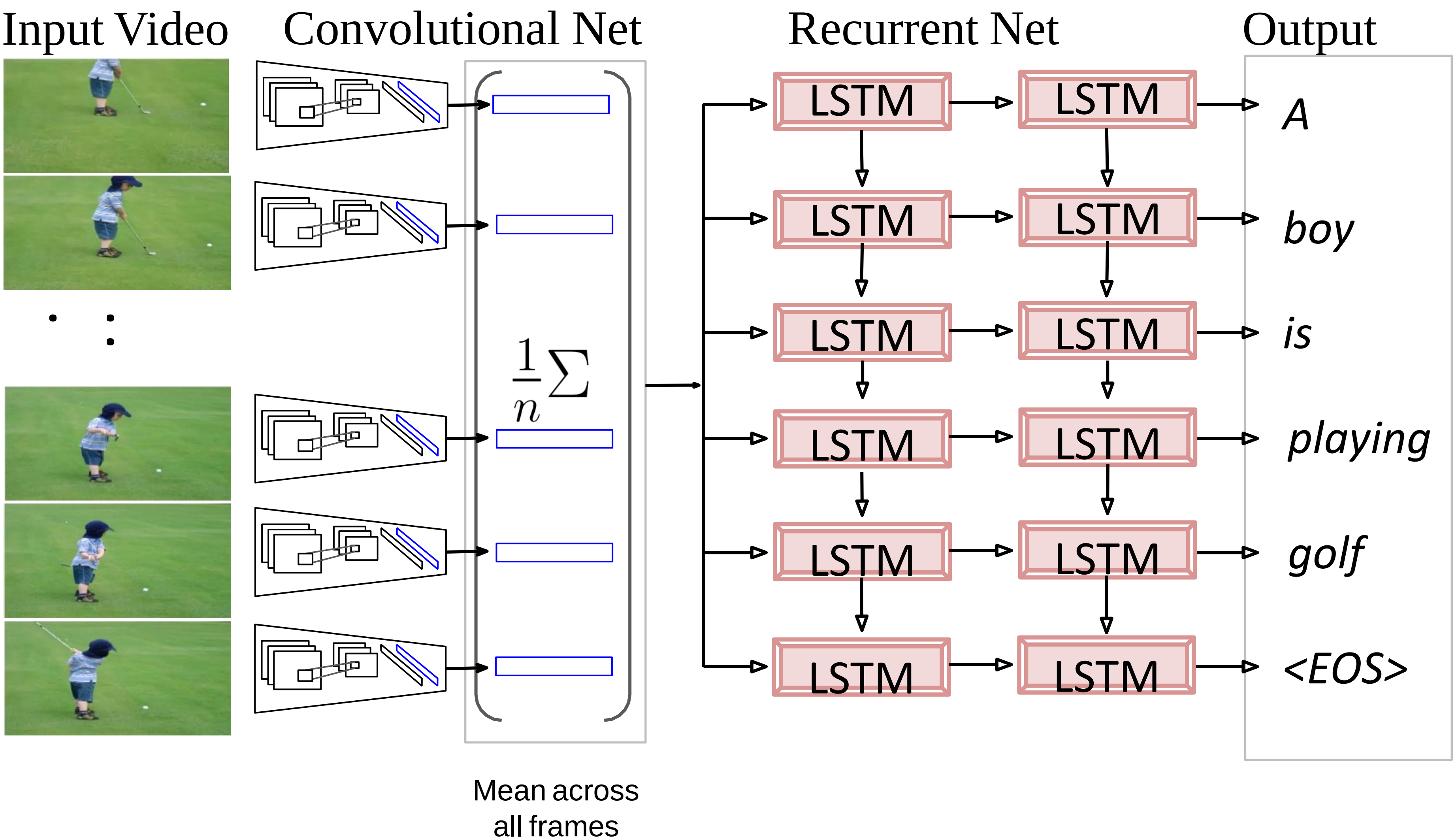
Input Video → Sample frames @1/10

Forward propagate
Output: “fc7” features
(activations before classification layer)

fc7: 4096 dimension
“feature vector”

Venugopalan et al., “Translating Videos to Natural Language using Deep Recurrent Neural Networks”, NAACL-HTL 2015

Video Captioning



Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning

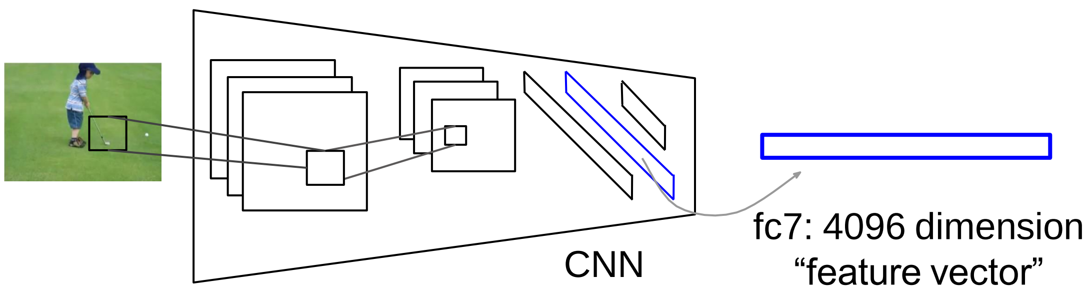
Annotated video data is scarce.

Key Insight:
Use supervised pre-training on data-rich
auxiliary tasks and transfer.

Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning

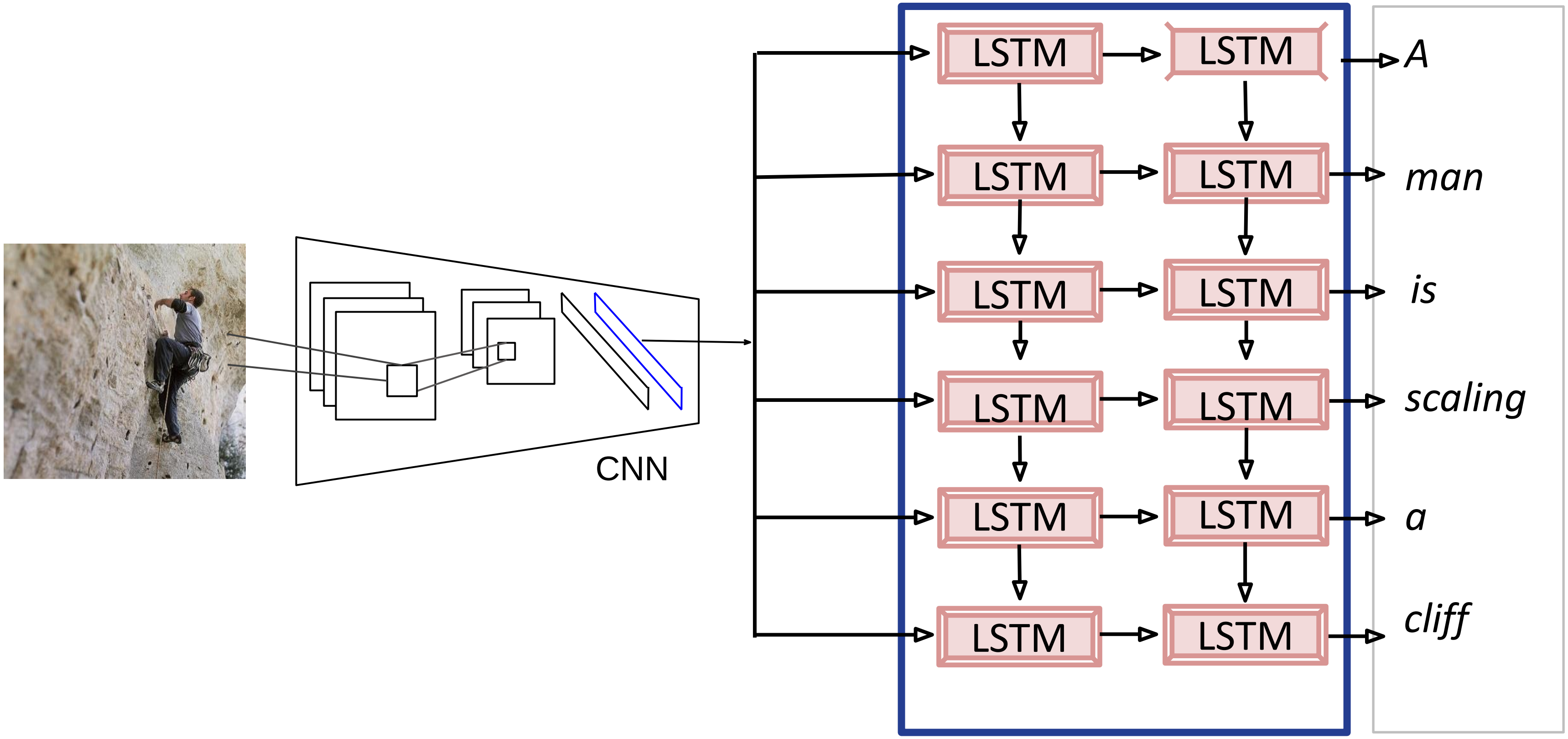
CNN pre-training



- Caffe Reference Net - variation of Alexnet [Krizhevsky et al. NIPS’12]
- 1.2M+ images from ImageNet ILSVRC-12 [Russakovsky et al.]
- Initialize weights of our network.

Video Captioning

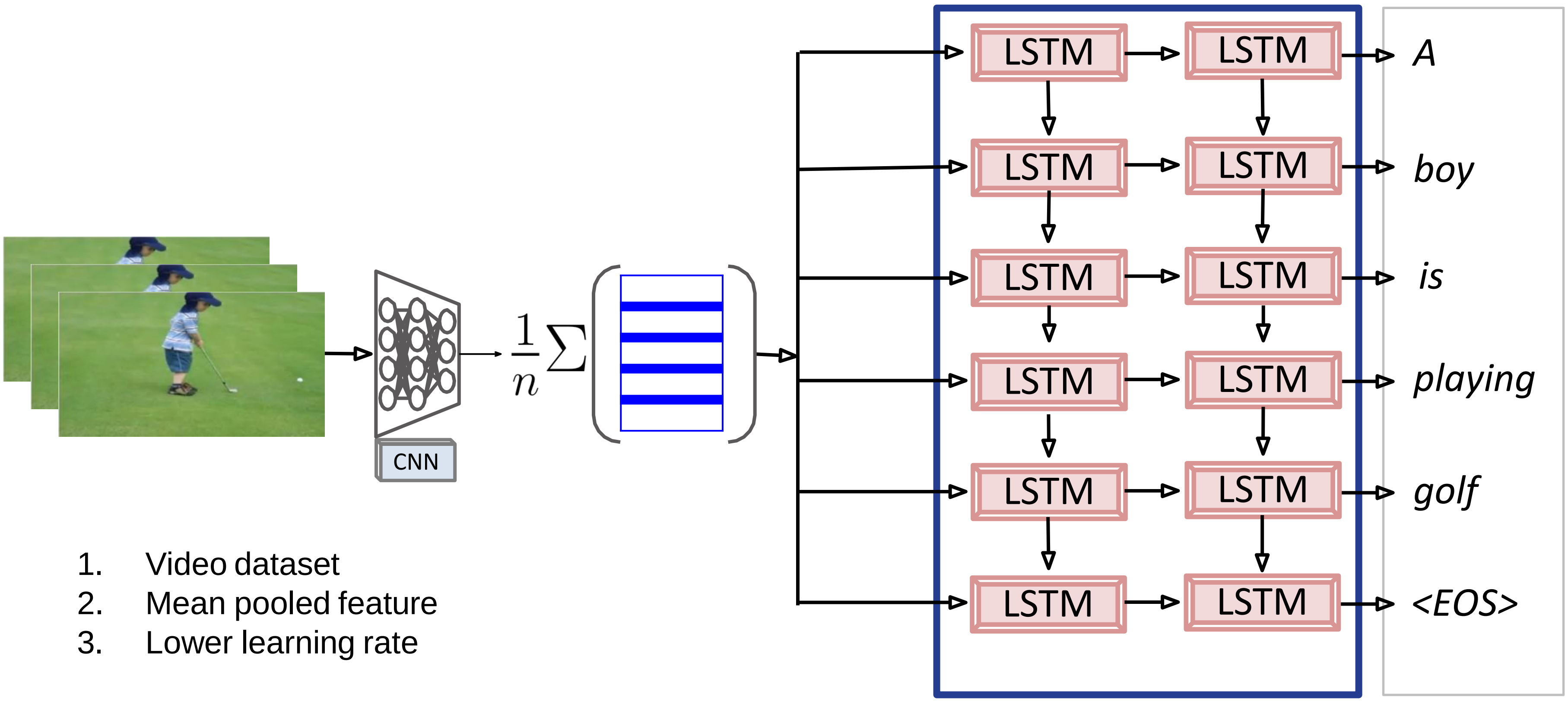
Image-Caption pre-training



Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning

Fine-tuning



1. Video dataset
2. Mean pooled feature
3. Lower learning rate

Venugopalan et al., “Translating Videos to Natural Language using Deep Recurrent Neural Networks”, NAACL-HTL 2015

Video Captioning

- A man appears to be plowing a rice field with a plow being pulled by two oxen.
- A man is plowing a mud field.
- Domesticated livestock are helping a man plow.
- A man leads a team of oxen down a muddy path.
- A man is plowing with some oxen.
- A man is tilling his land with an ox pulled plow.
- Bulls are pulling an object.
- Two oxen are plowing a field.
- The farmer is tilling the soil.
- A man in ploughing the field.

- A man is walking on a rope.
- A man is walking across a rope.
- A man is balancing on a rope.
- A man is balancing on a rope at the beach.
- A man walks on a tightrope at the beach.
- A man is balancing on a volleyball net.
- A man is walking on a rope held by poles
- A man balanced on a wire.
- The man is balancing on the wire.
- A man is walking on a rope.
- A man is standing in the sea shore.

Venugopalan et al., “Translating Videos to Natural Language using Deep Recurrent Neural Networks”, NAACL-HTL 2015

Video Captioning

MT metrics (BLEU, METEOR) to compare the system generated sentences against (all) ground truth references.

Model	BLEU	METEOR
Best Prior Work [Thomason et al. COLING'14]	13.68	23.90
Only Images	12.66	20.96
Only Video	31.19	26.87
Images+Video	**33.29**	**29.07**

Pre-training only, no fine-tuning

No pre-training

Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning

FGM: A person is dancing with the person on the stage.
YT: A group of men are riding the forest.
I+V: **A group of people are dancing.**
GT: Many men and women are dancing in the street.

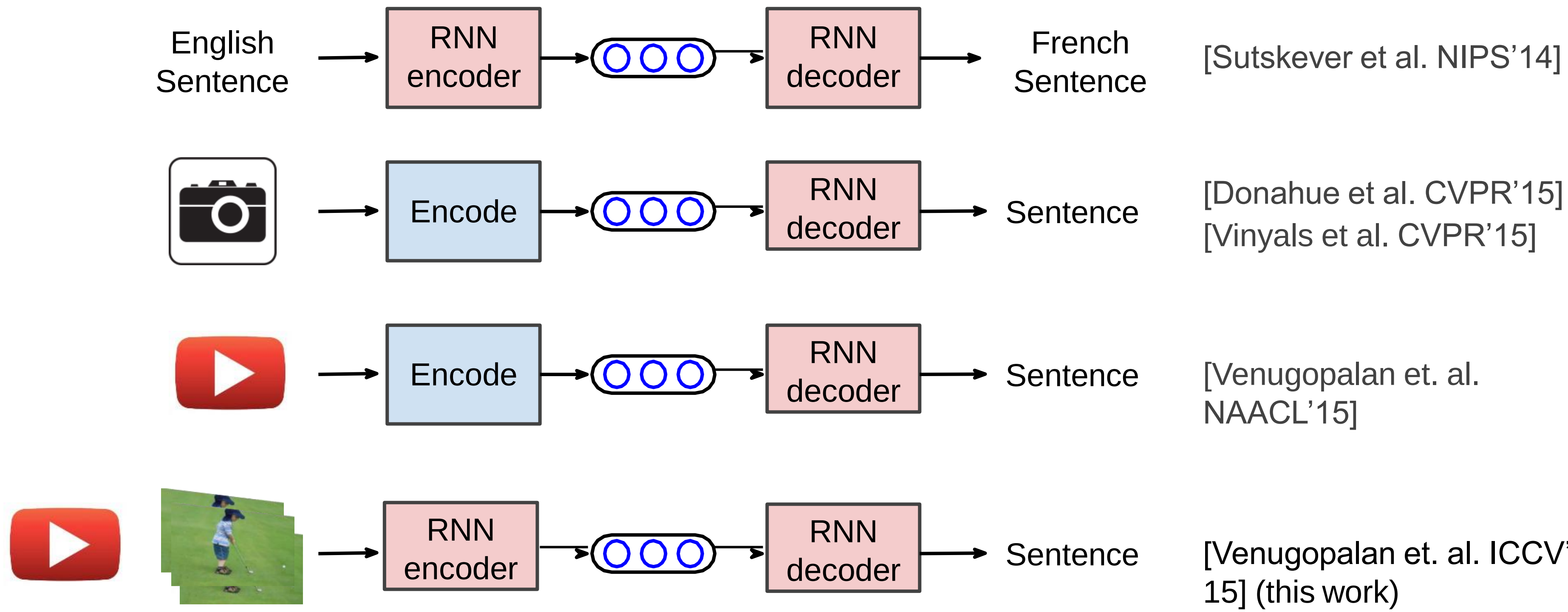
FGM: A person is cutting a potato in the kitchen.
YT: A man is slicing a tomato.
I+V: **A man is slicing a carrot.**
GT: A man is slicing carrots.

FGM: A person is walking with a person in the forest.
YT: A monkey is walking.
I+V: **A bear is eating a tree.**
GT: Two bear cubs are digging into dirt and plant matter at the base of a tree.

FGM: A person is riding a horse on the stage.
YT: A group of playing are playing in the ball.
I+V: **A basketball player is playing.**
GT: Dwayne wade does a fancy layup in an allstar game.

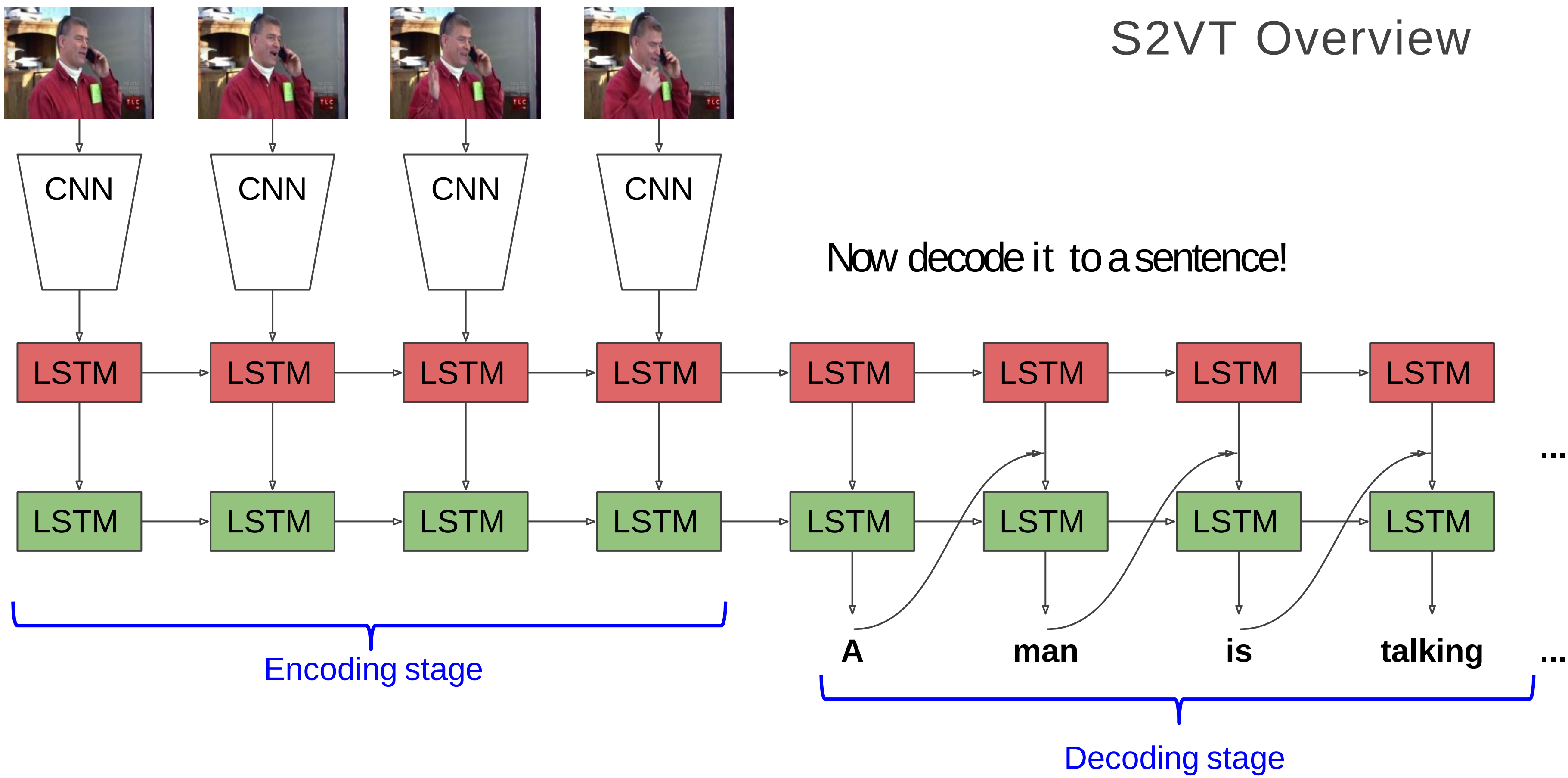
Venugopalan et al., "Translating Videos to Natural Language using Deep Recurrent Neural Networks", NAACL-HTL 2015

Video Captioning



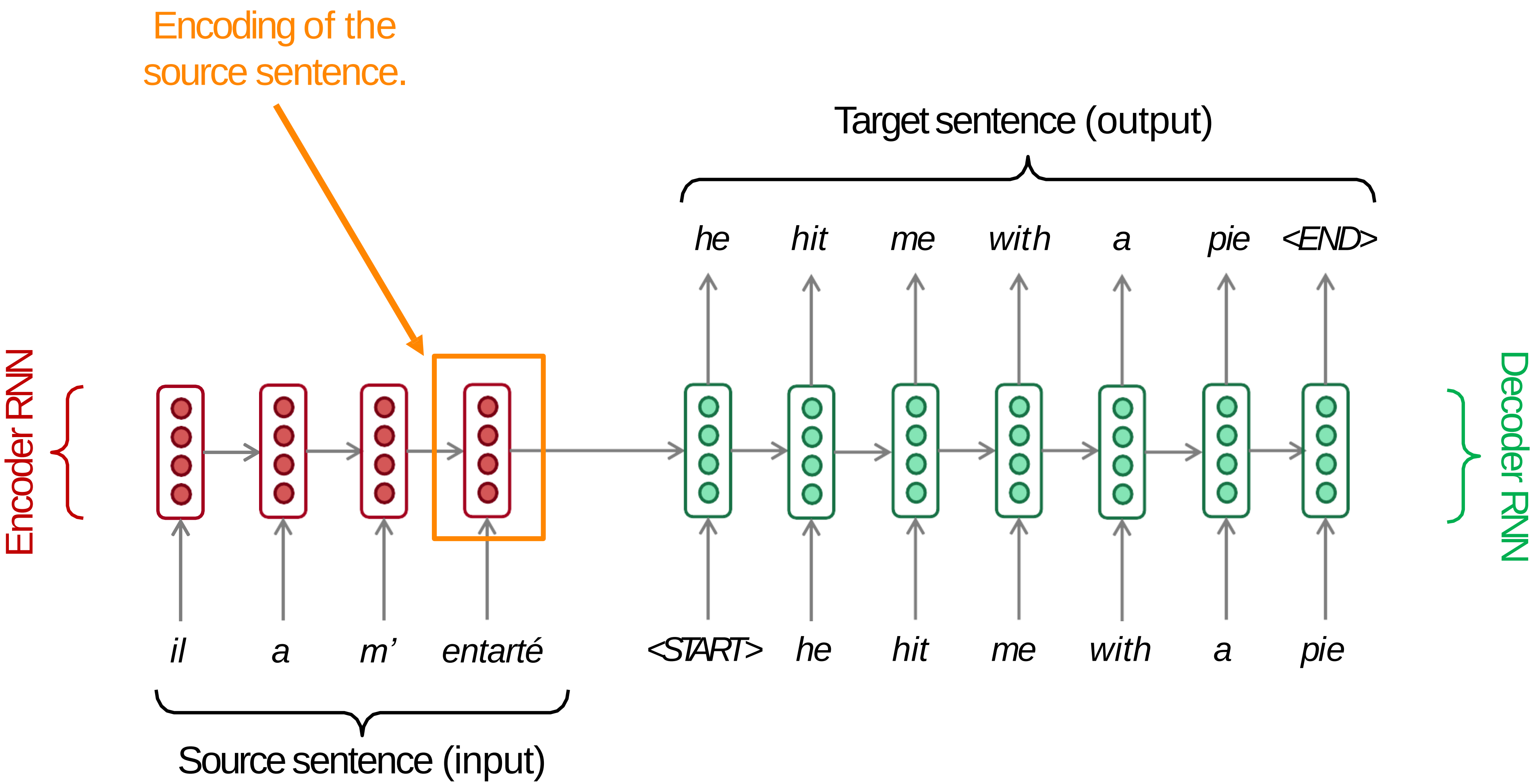
Venugopalan et al., "Sequence to Sequence - Video to Text", ICCV 2015

Video Captioning



Venugopalan et al., “Sequence to Sequence - Video to Text”, ICCV 2015

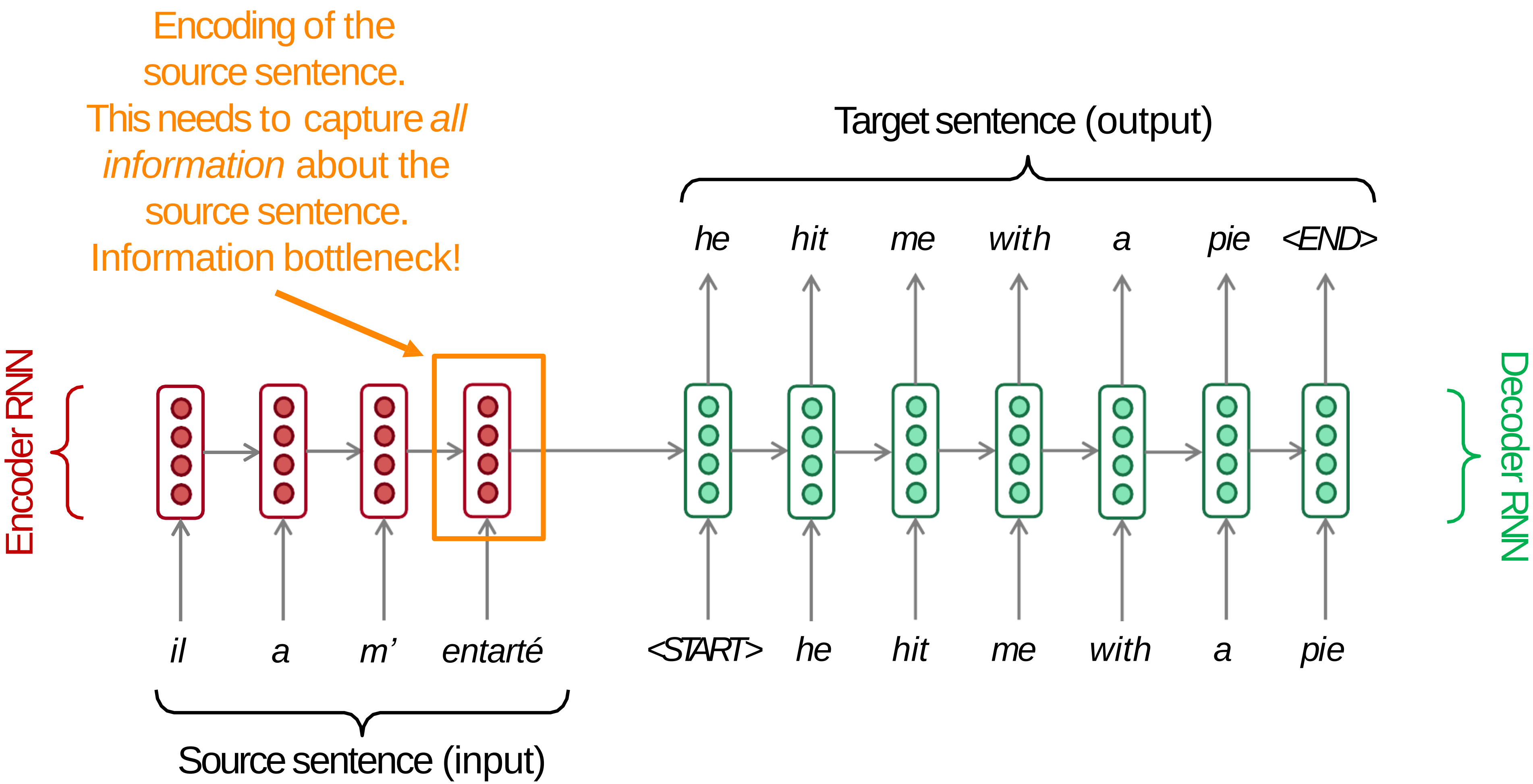
Sequence-to-sequence: the bottleneck problem



Problems with this architecture?

Abigail See

Sequence-to-sequence: the bottleneck problem



Abigail See

Attention

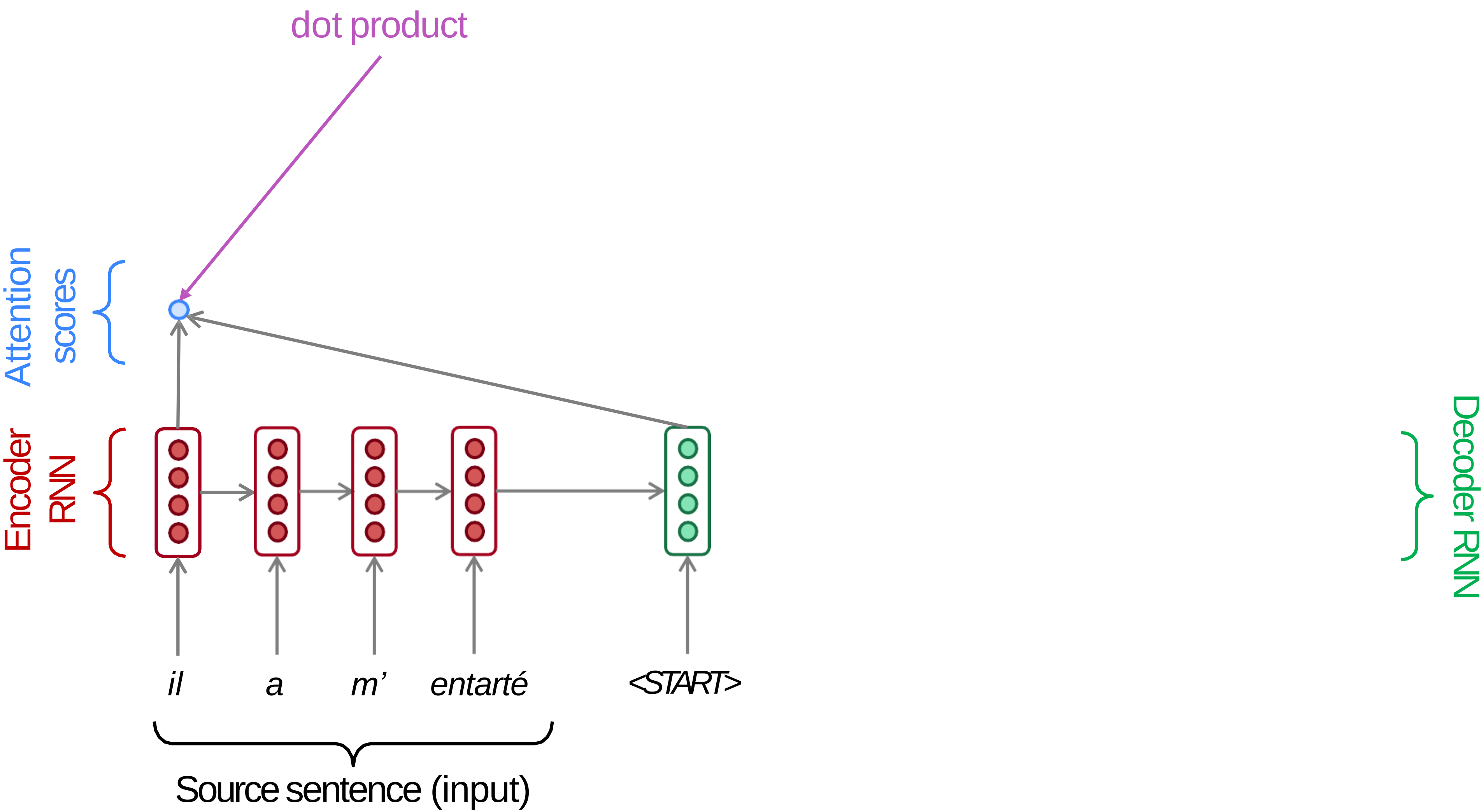
- **Attention** provides a solution to the bottleneck problem.

- <u>Core idea</u>: on each step of the decoder, use *direct connection to the encoder* to *focus on a particular part* of the source sequence

- First we will show via diagram (no equations), then we will show with equations

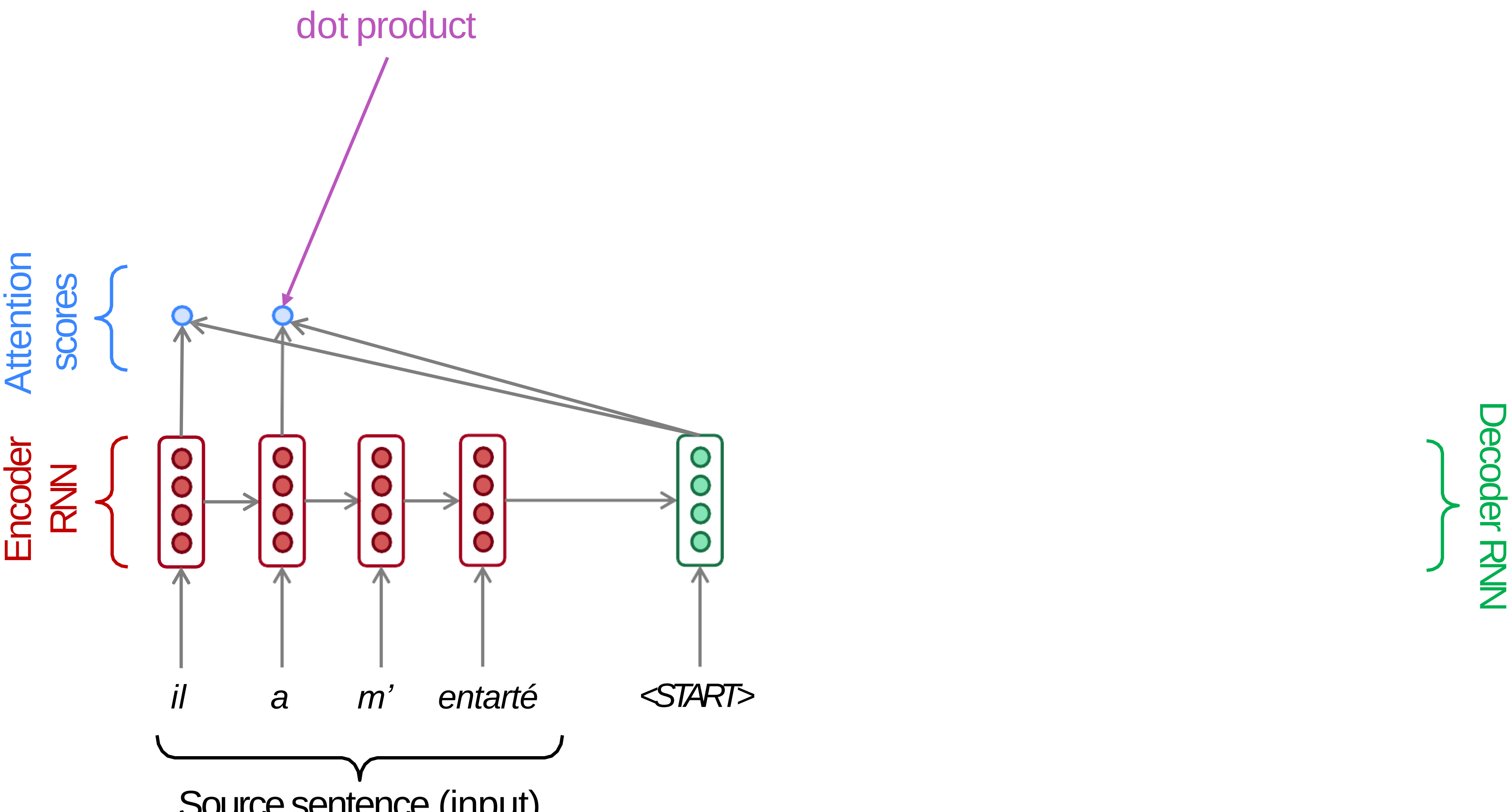
Abigail See

Sequence-to-sequence with attention



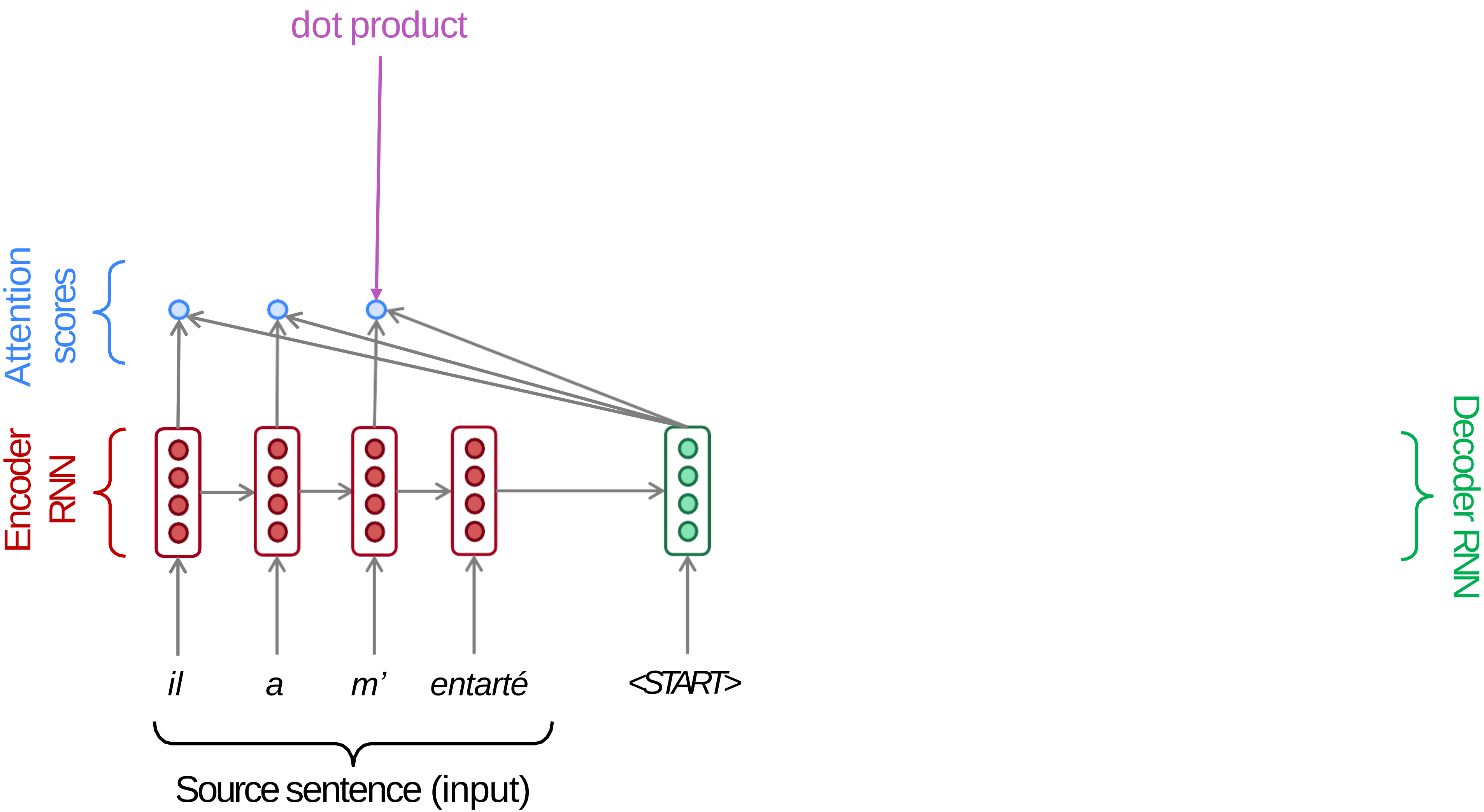
Abigail See

Sequence-to-sequence with attention



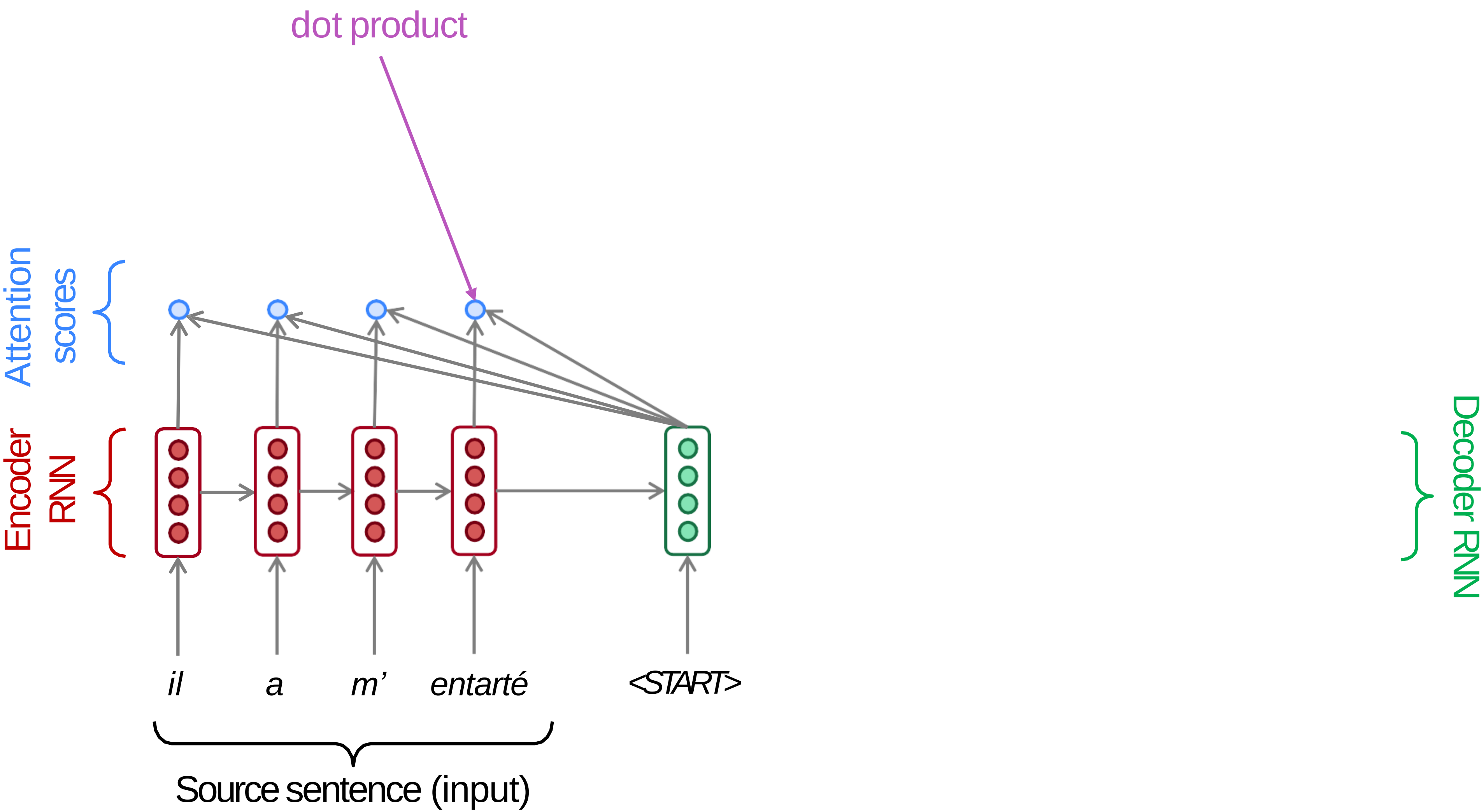
Abigail See

Sequence-to-sequence with attention



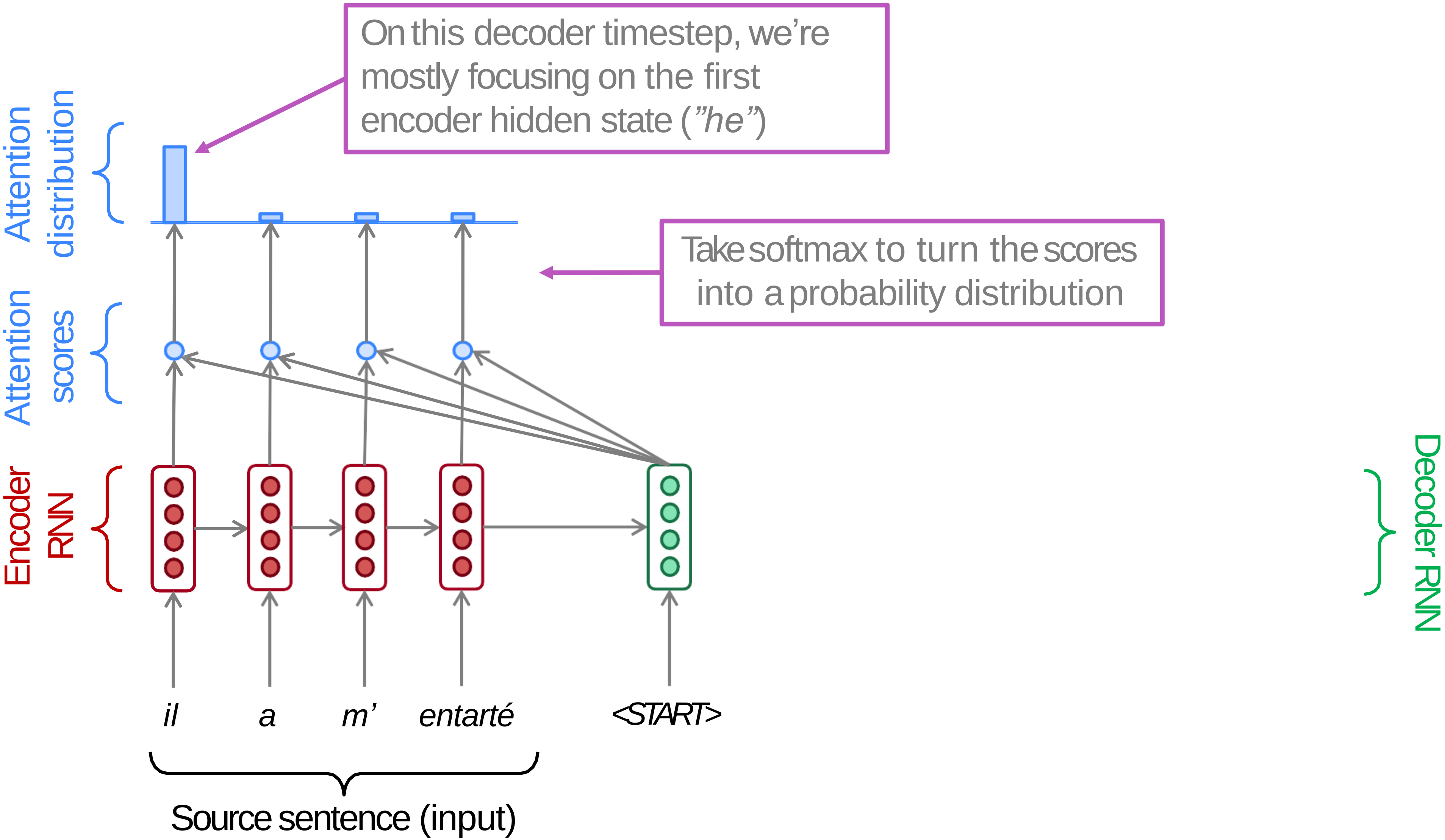
Abigail See

Sequence-to-sequence with attention



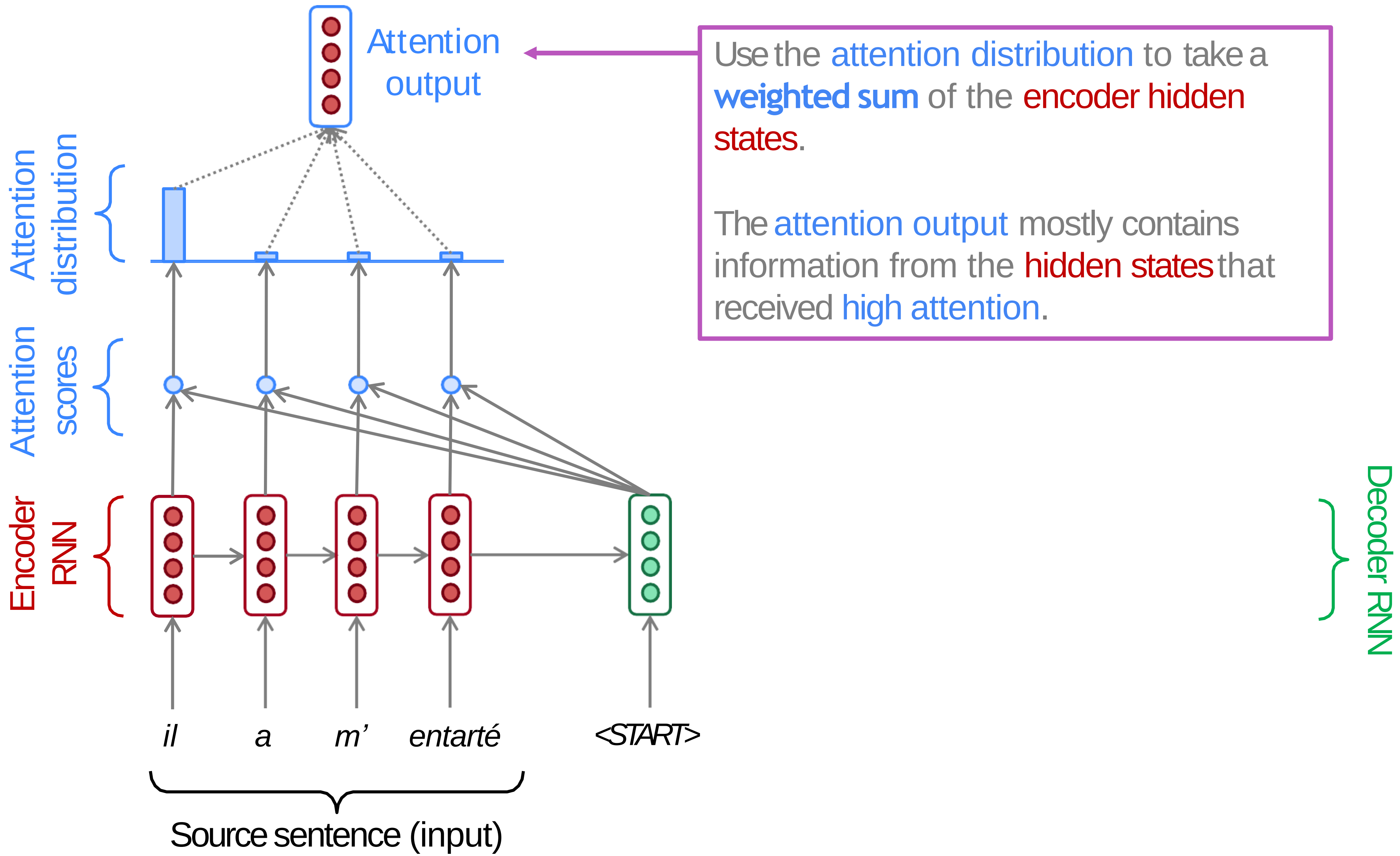
Abigail See

Sequence-to-sequence with attention



Abigail See

Sequence-to-sequence with attention

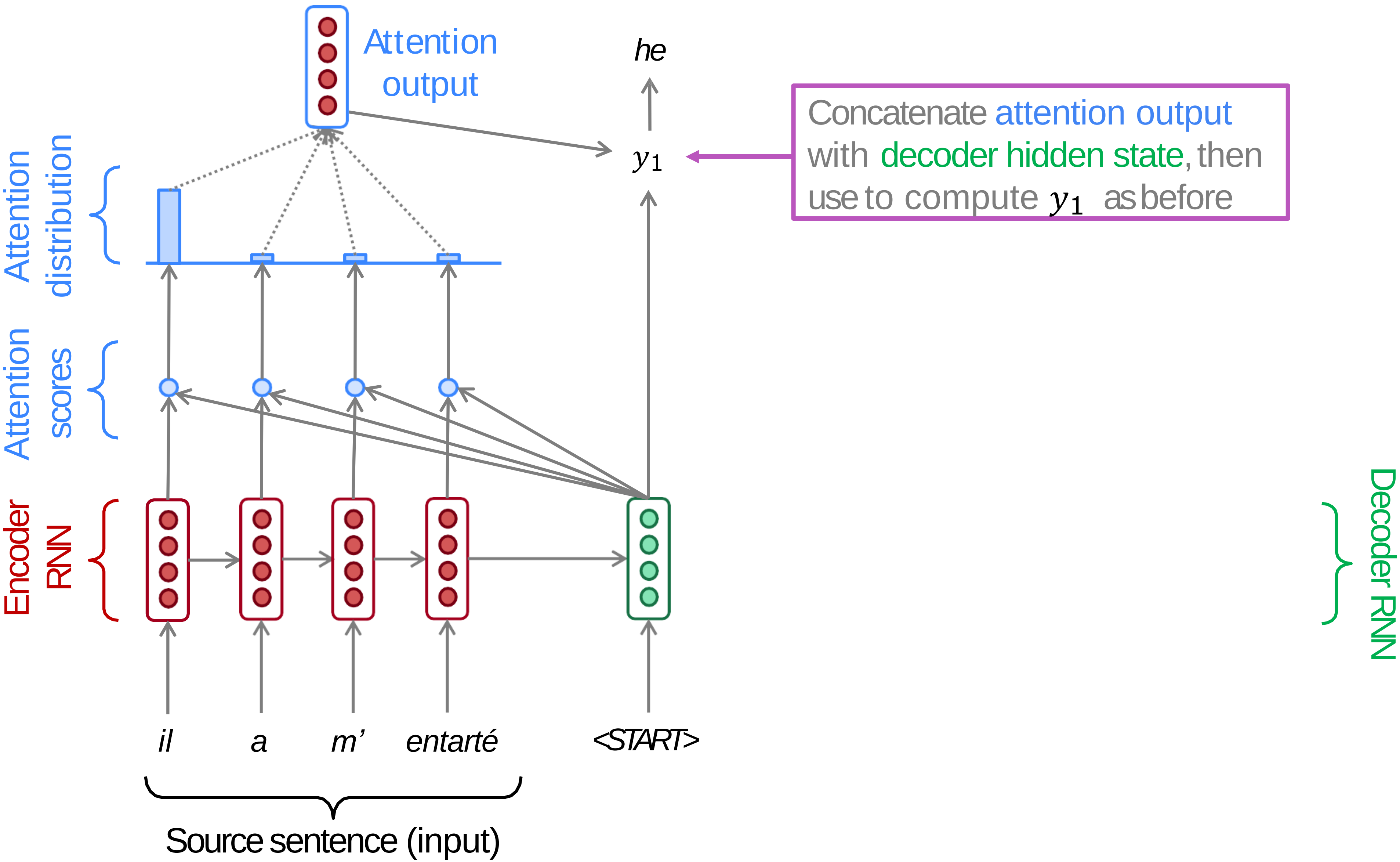


Use the attention distribution to take a **weighted sum** of the encoder hidden states.

The attention output mostly contains information from the hidden states that received high attention.

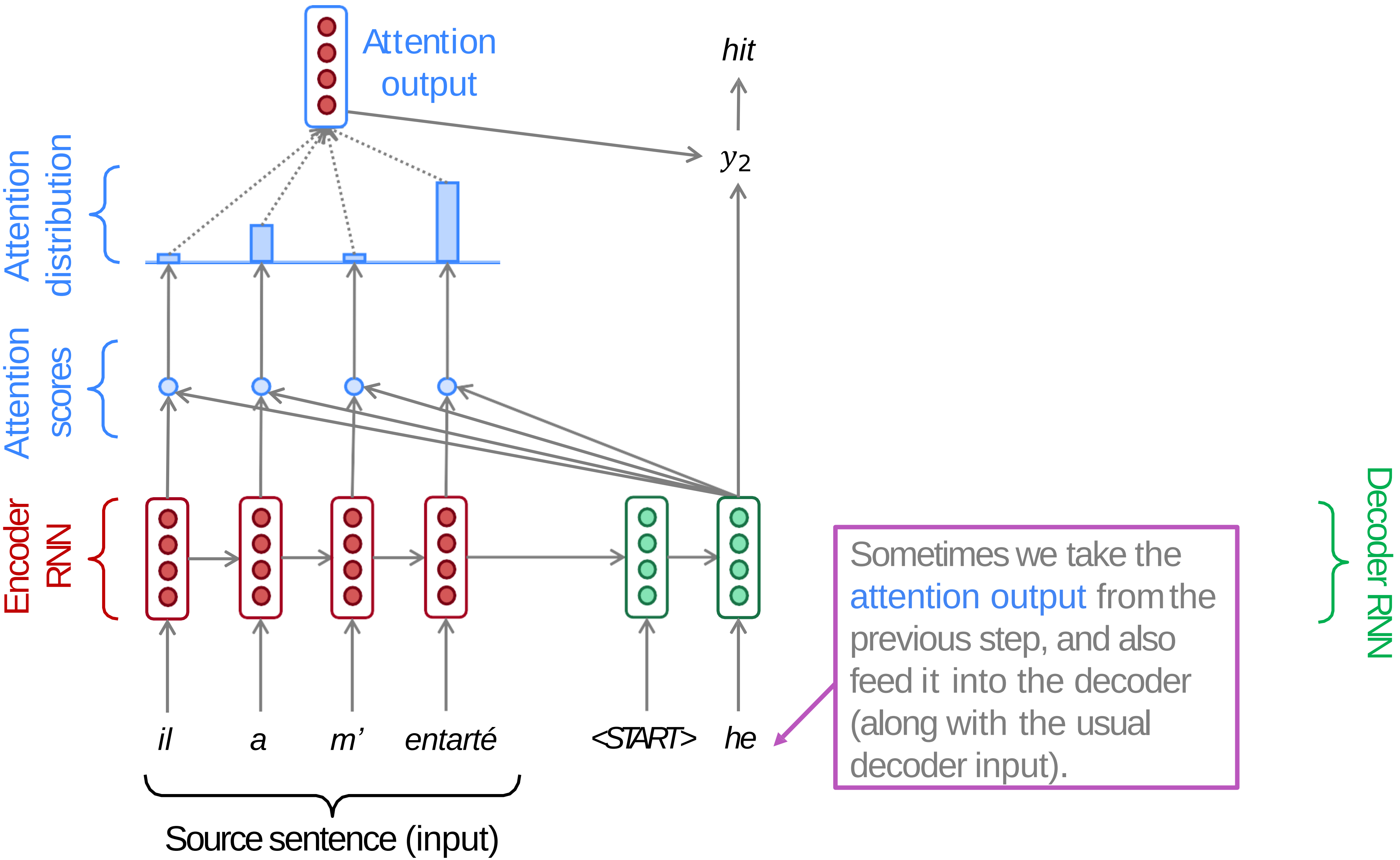
Abigail See

Sequence-to-sequence with attention



Abigail See

Sequence-to-sequence with attention



Sometimes we take the attention output from the previous step, and also feed it into the decoder (along with the usual decoder input).

Abigail See

Attention: in equations

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$
- We get the attention score $\boldsymbol{e}^t$ for this step:

$$\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$$

- We take softmax to get the attention distribution α^t for this step (this is a probability distribution and sums to 1)

$$\alpha^t = \mathrm{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

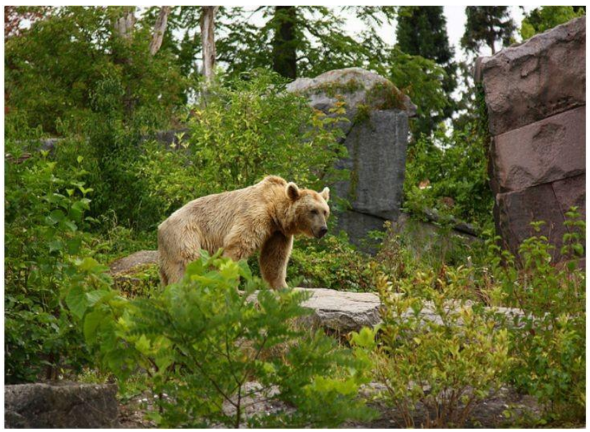
- We use α^t to take a weighted sum of the encoder hidden states to get the attention output $\boldsymbol{a}_t$

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

- Finally we concatenate the attention output $\boldsymbol{a}_t$ with the decoder hidden state s_t and proceed as in the non-attention seq2seq model

$$[\boldsymbol{a}_t; \boldsymbol{s}_t] \in \mathbb{R}^{2h}$$

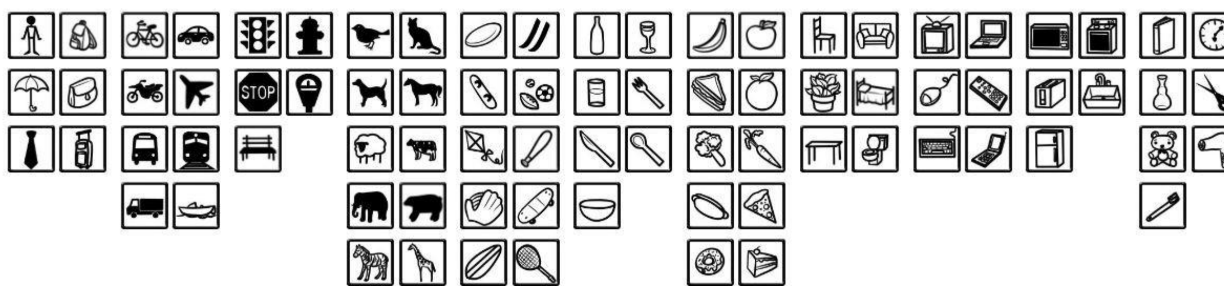
Visual Description



Berkeley LRCN [Donahue et al. CVPR'15]:
A brown bear standing on top of a lush green field.

MSR CaptionBot [http://captionbot.ai/]:
A large brown bear walking through a forest.

MSCOCO
80 classes

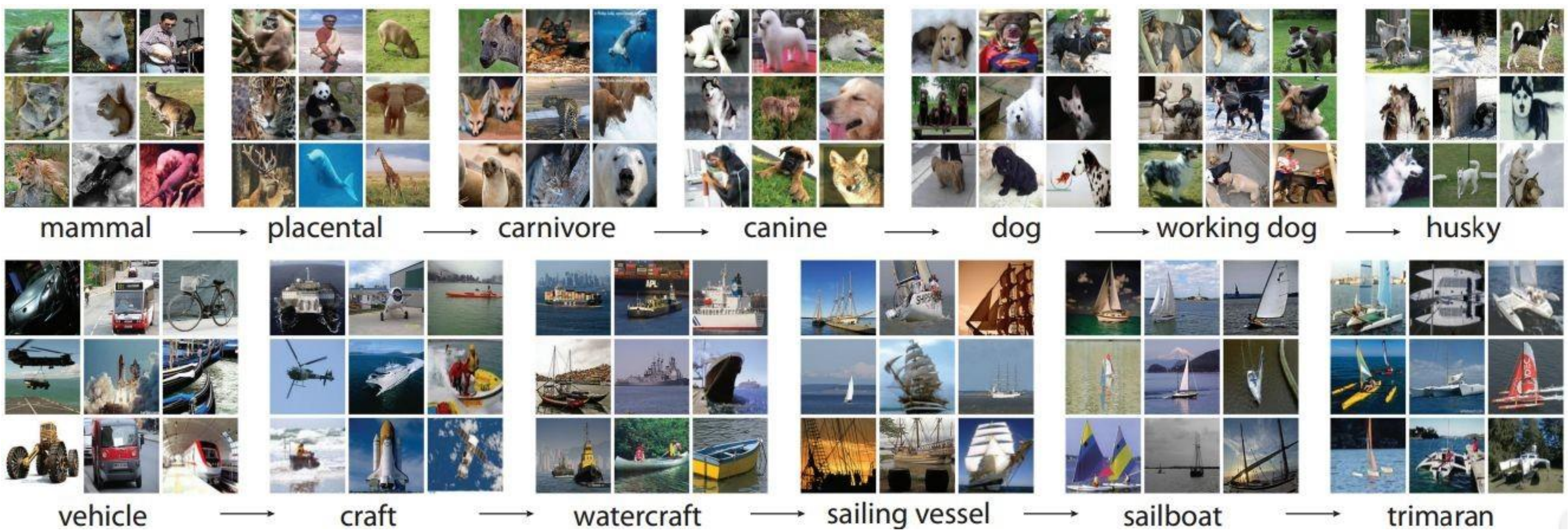


Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Object Recognition

Can identify hundreds of categories of objects.

IMAGENET 14M images, 22K classes [Deng et al. CVPR'09]



Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Novel Object Captioner (NOC)

compose descriptions of 100s of objects in context

NOC (ours): Describe novel objects <u>without</u> paired image-caption data.

IMAGENET + MSCOCO +

→ An **okapi** standing in the middle of a field.

Visual Classifiers. IMAGENET

→ **okapi**

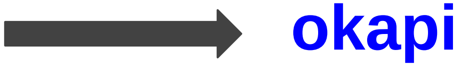
Existing captioners. IMAGENET init + train MSCOCO

→ A horse standing in the dirt.

Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Insights

1. Need to recognize and describe objects outside of image-caption datasets.



Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Insights

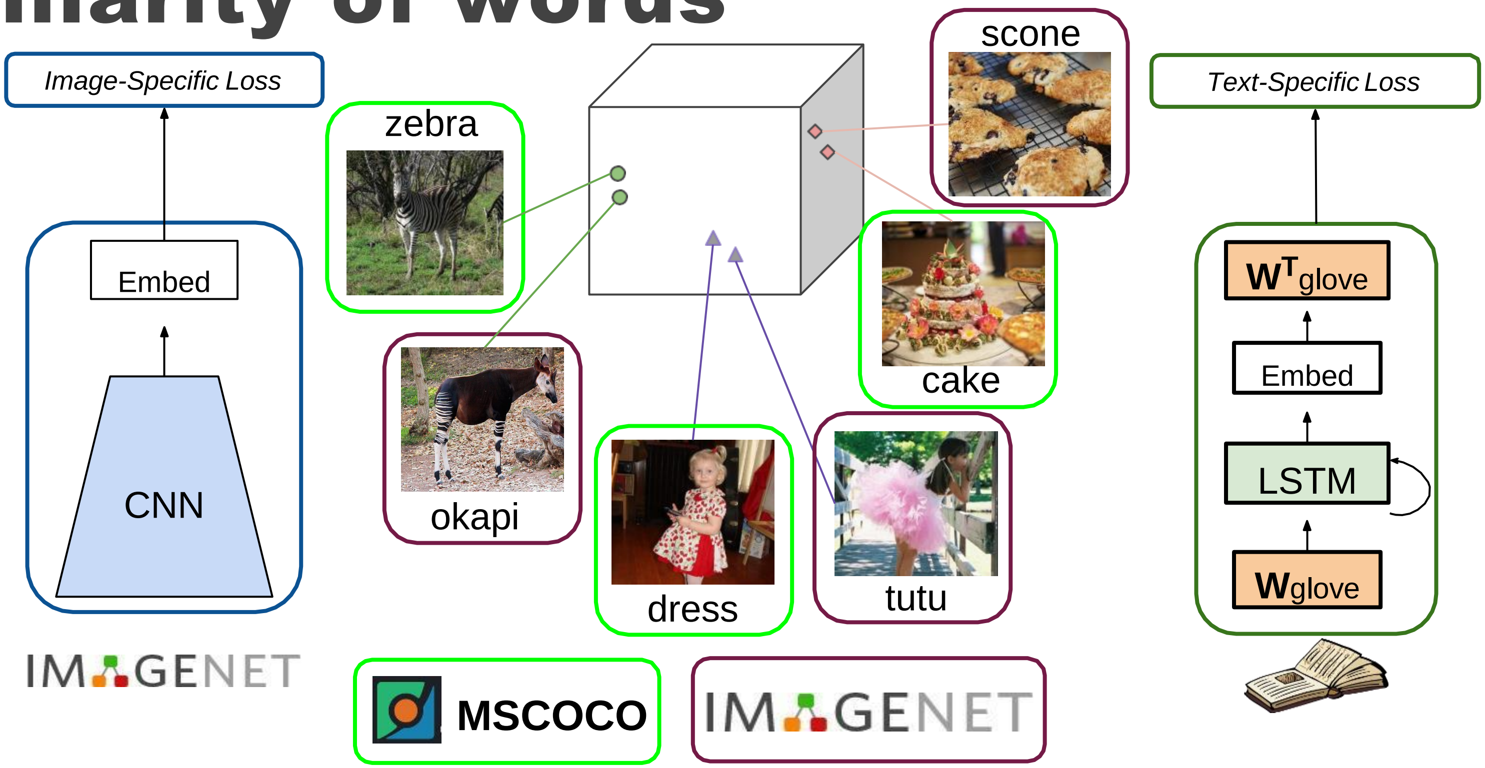
2. Describe unseen objects that are similar to objects seen in image-caption datasets.

okapi

zebra

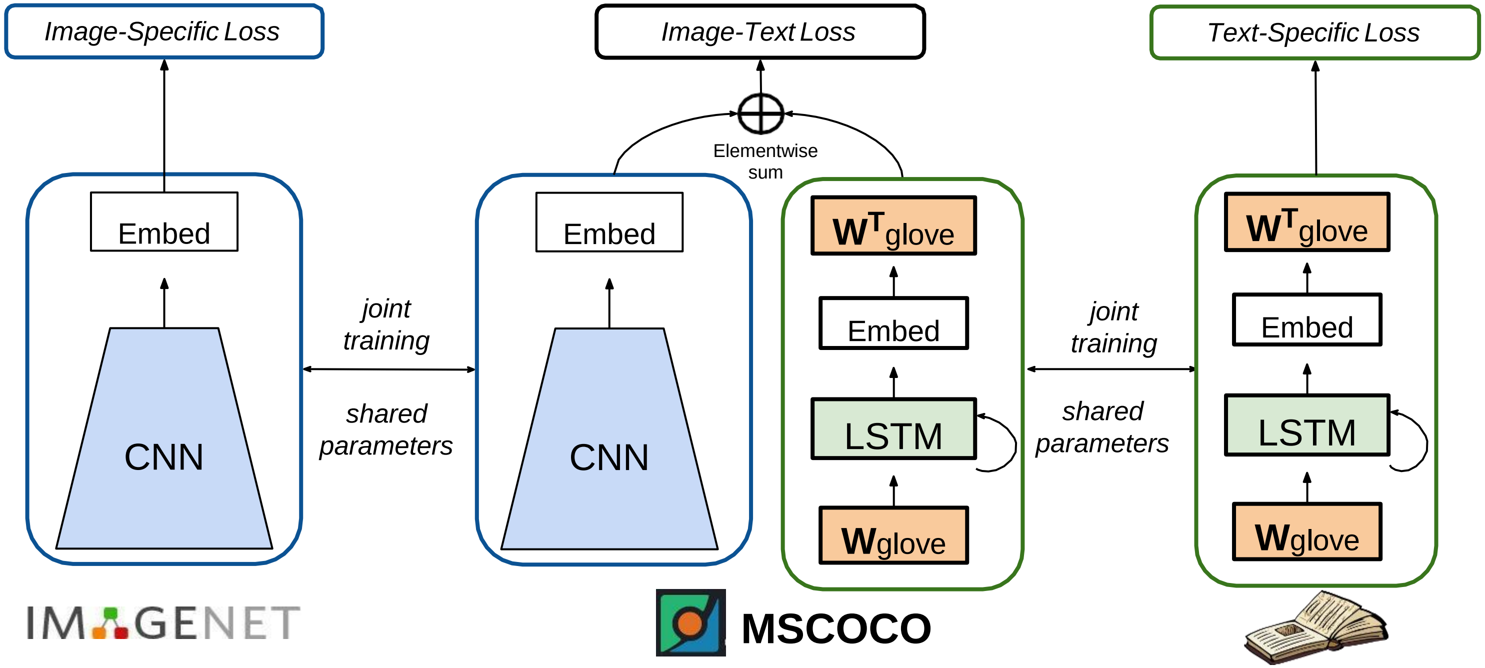
Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Insight 2: Capture semantic similarity of words



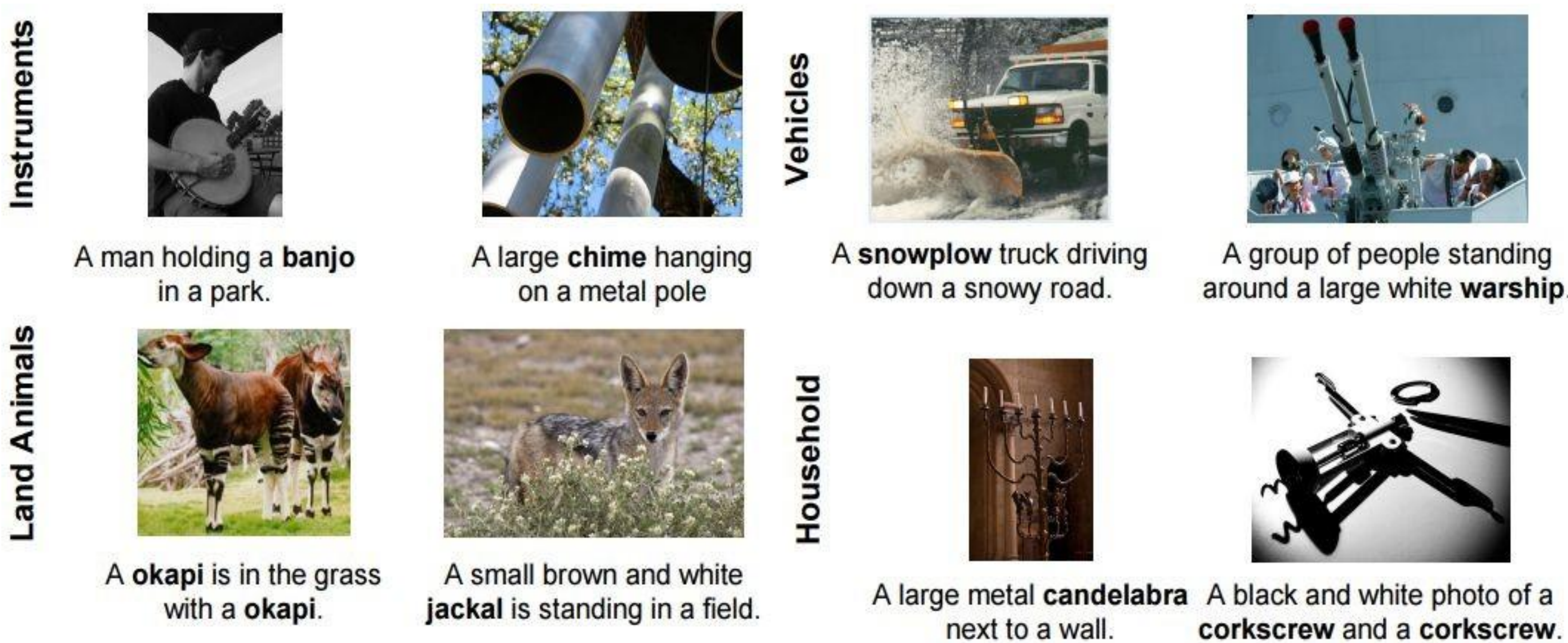
Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Insight 3: Jointly train on multiple sources



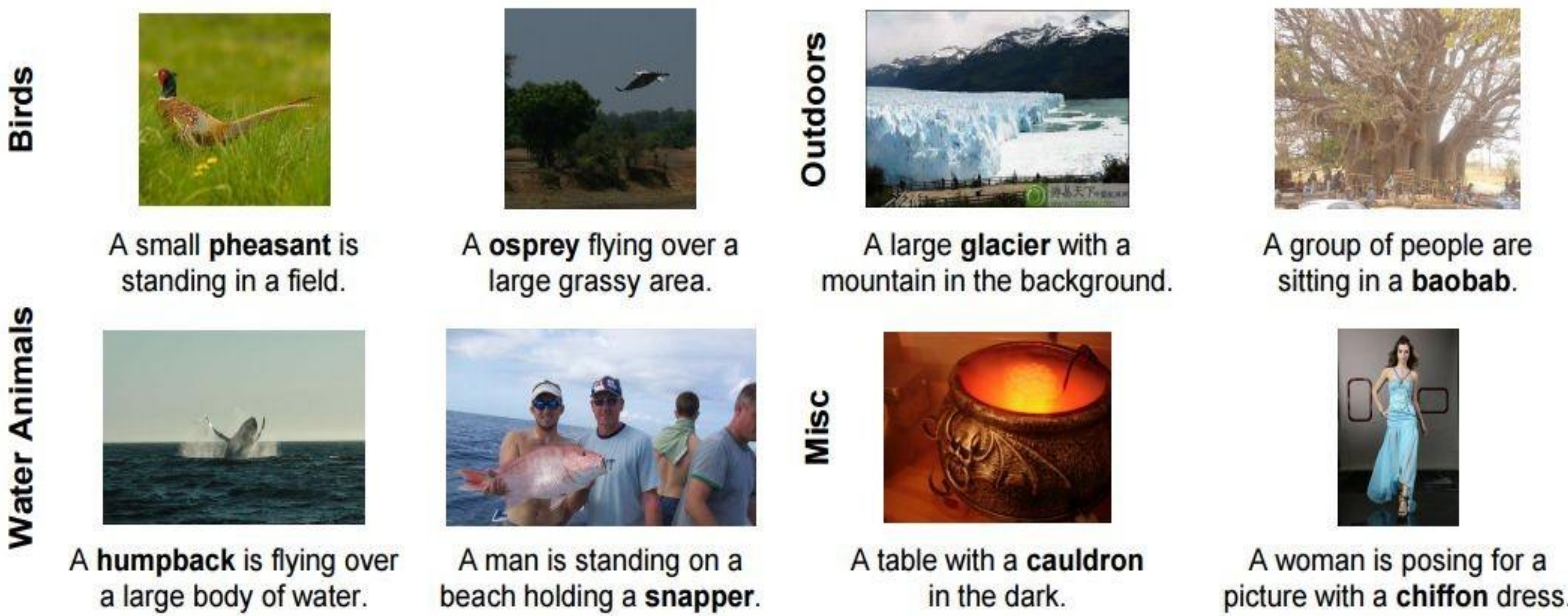
Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Qualitative Evaluation: ImageNet



Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Qualitative Evaluation: ImageNet



Venugopalan et al., "Captioning Images With Diverse Objects", CVPR 2017

Plan for this lecture

- Learning the relation between images and text
  - Recurrent neural networks
  - Applications: Captioning
  - Transformers
- Reasoning: Visual question answering
  - Neuro-symbolic VQA
  - Graph convolutional networks
- Multimodal self-supervised learning

Transformers: Motivation

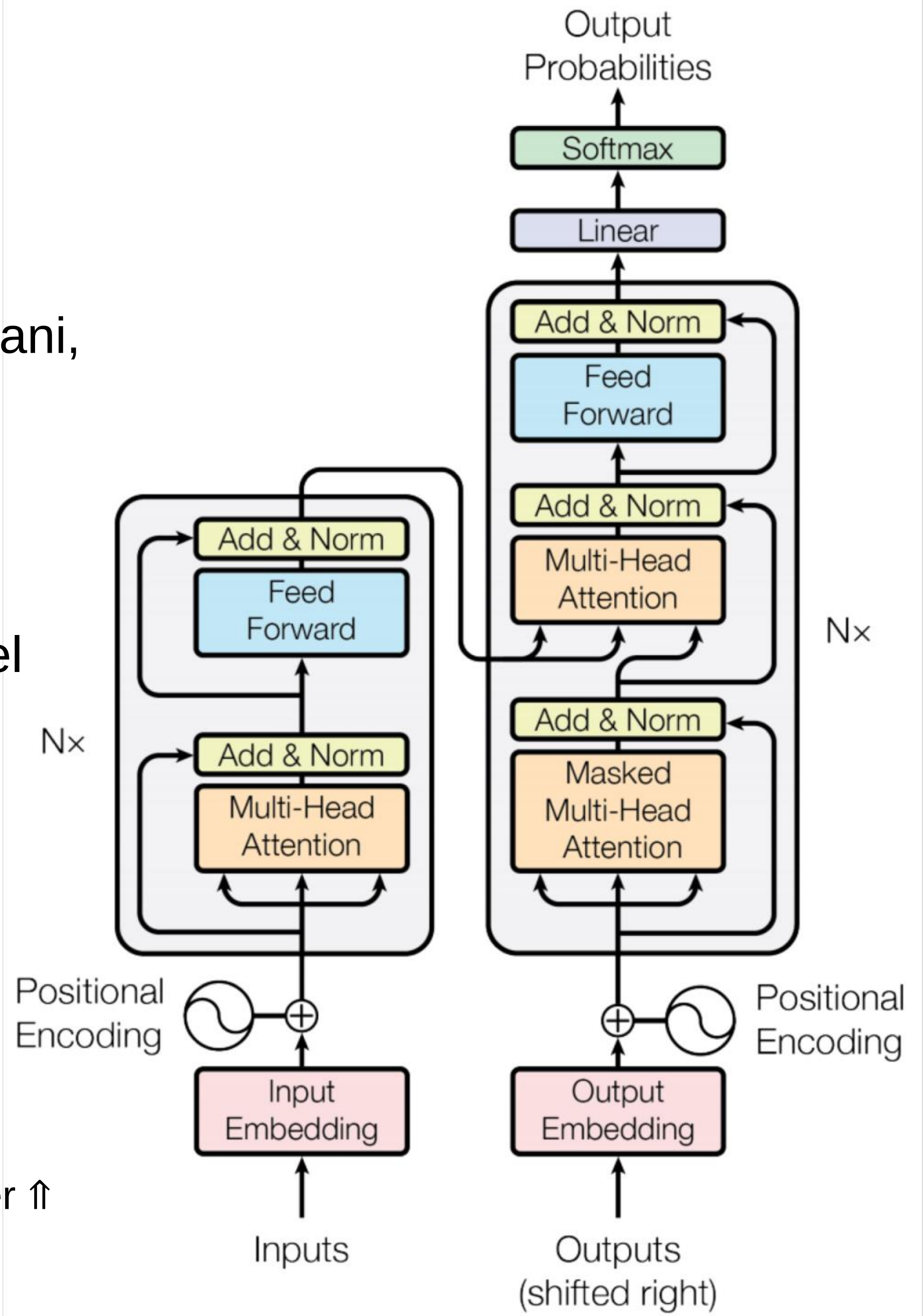
- We want **parallelization** but RNNs are inherently sequential
- Despite GRUs and LSTMs, RNNs still need attention mechanism to deal with long range dependencies – **path length** between states grows with sequence otherwise
- But if **attention** gives us access to any state… maybe we can just use attention and don't need the RNN?

Adapted from Christopher Manning

Transformer Overview

Attention is all you need. 2017. Aswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin
https://arxiv.org/pdf/1706.03762.pdf

- Non-recurrent sequence-to-sequence encoder-decoder model
- Task: machine translation with parallel corpus
- Predict each translated word
- Final cost/error function is standard cross-entropy error on top of a softmax classifier



This and related figures from paper ⇑

Dot-Product Attention (Extending our previous def.)

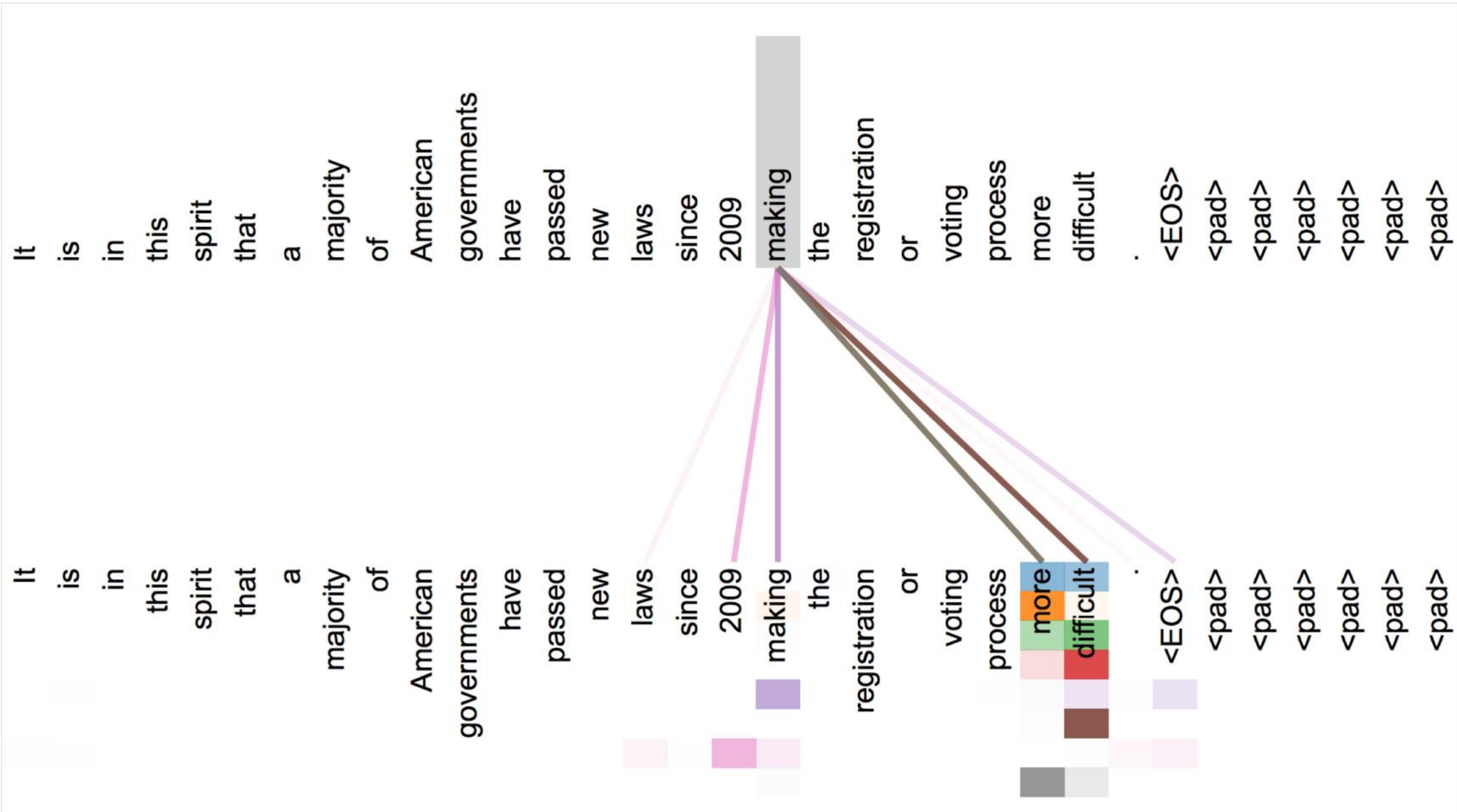
- Inputs: query q and set of key-value (k-v) pairs to an output
- Query, keys, values, and output are all vectors

- Output is weighted sum of values, where
- Weight of each value is computed by an inner product of query and corresponding key
- Queries, keys have same dimensionality d_k , value have d_v

$$A(q, K, V) = \sum_i \frac{e^{q \cdot k_i}}{\sum_j e^{q \cdot k_j}} v_i$$

Attention visualization in layer 5

- Words start to pay attention to other words in sensible ways



https://github.com/jessevig/bertviz

BERT: Devlin, Chang, Lee, Toutanova (2018)

- Mask out *k*% of the input words, and then predict the masked words
  - They always use *k* = 15%

store gallon

↑ ↑

the man went to the [MASK] to buy a [MASK] of milk

- Too little masking: Too expensive to train
- Too much masking: Not enough context

Additional task: Next sentence prediction

- To learn *relationships* between sentences, predict whether Sentence B is actual sentence that follows Sentence A, or a random sentence

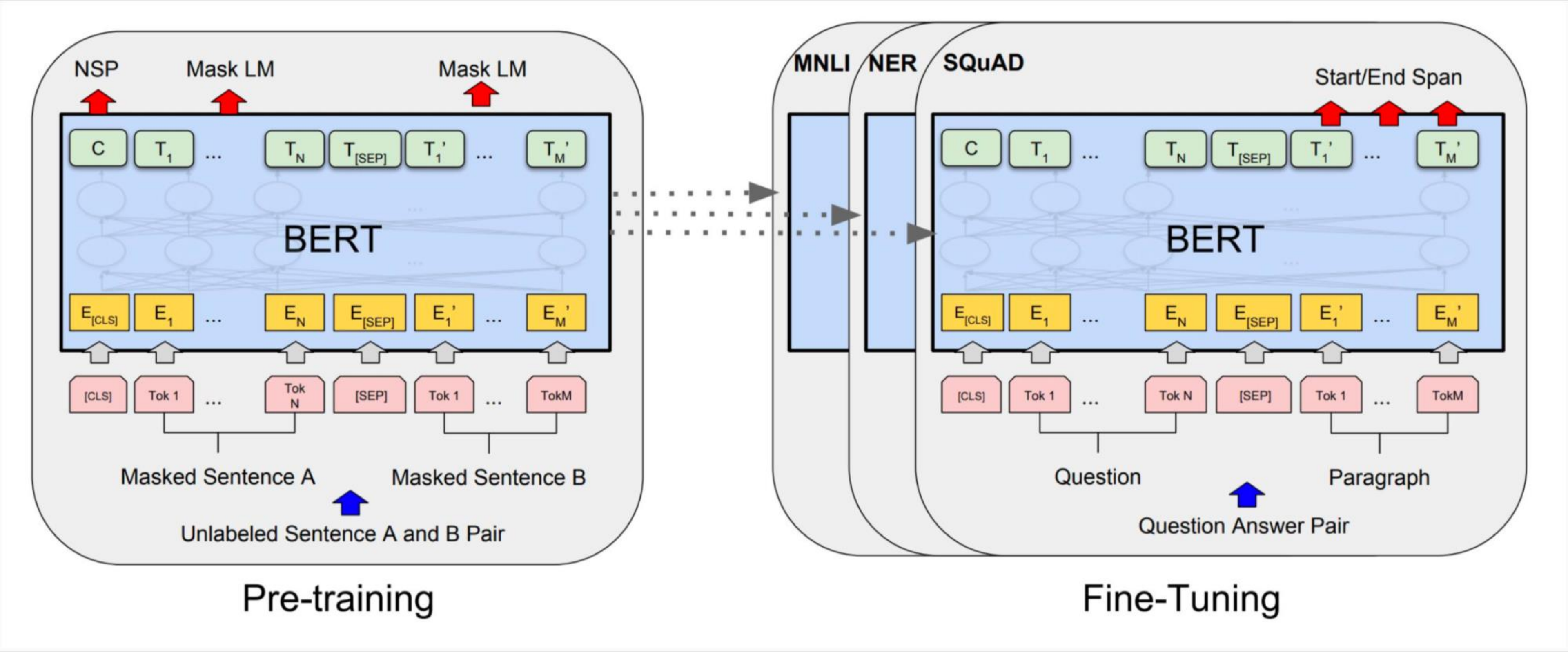
Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Adapted from Christopher Manning

BERT model fine tuning

- Simply learn a classifier built on the top layer for each task that you fine tune for



SQuAD 2.0 leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Cross-modal transformers

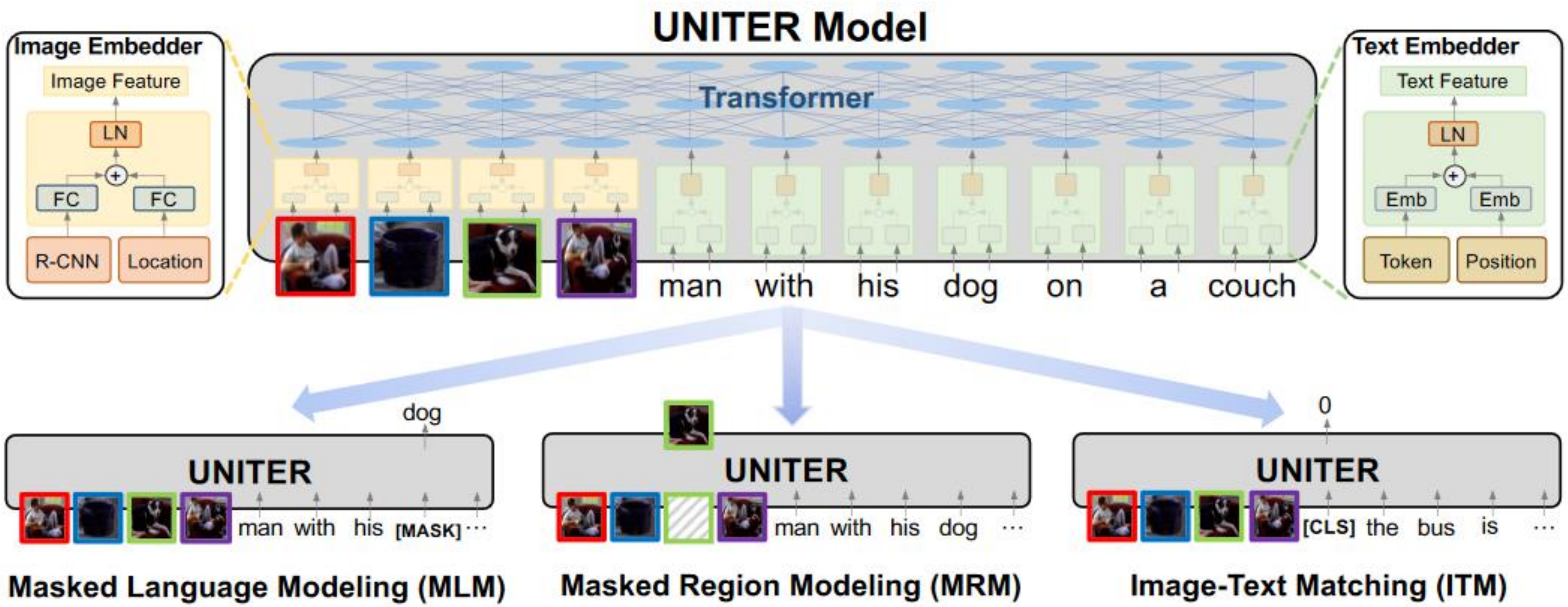


Figure 1: Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer self-attention Transformer, learned through three pre-training tasks.

Chen et al., "UNITER: Learning UNiversal Image-TExt Representations", arxiv 2019

Cross-modal transformers

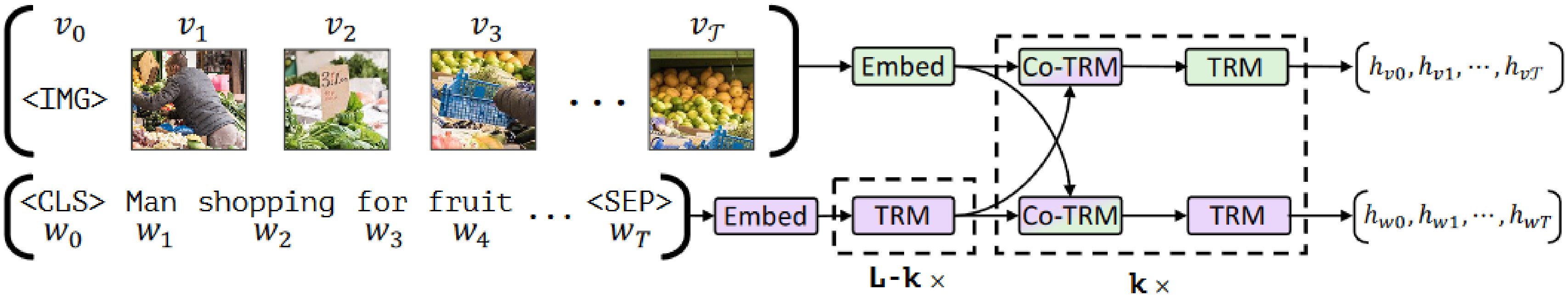


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

Lu et al., “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”, NeurIPS 2019

Cross-modal transformers

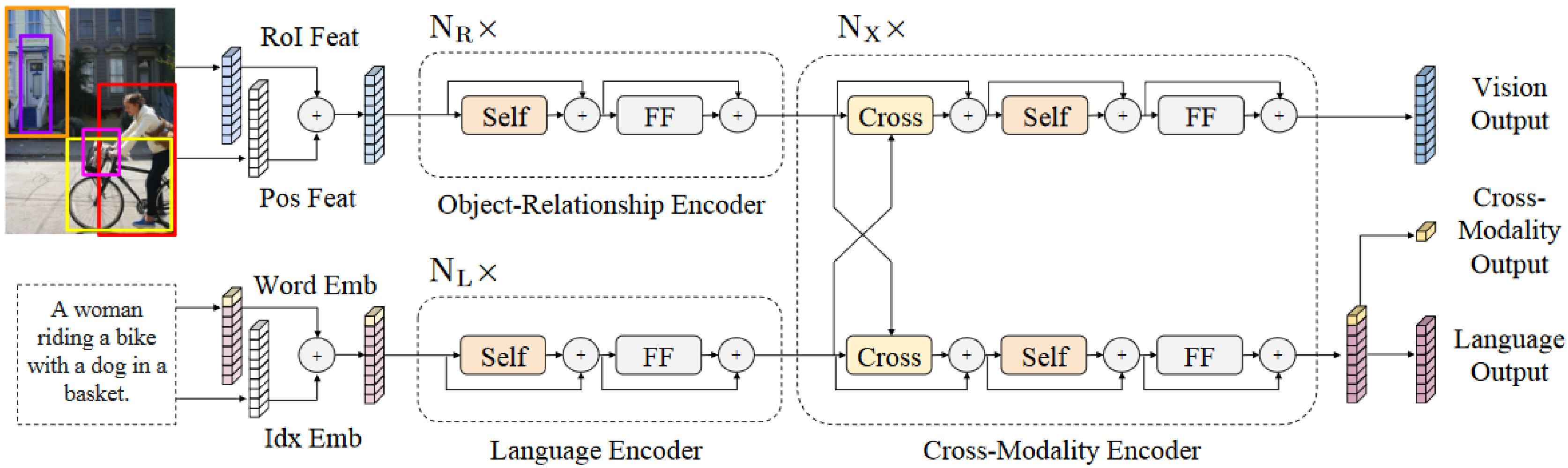
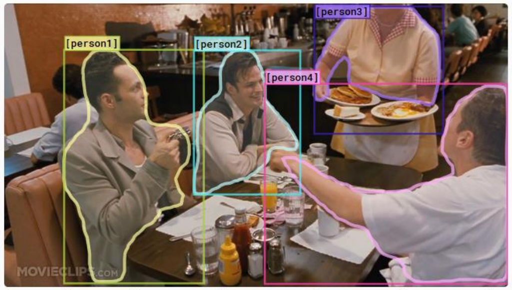


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

Tan and Bansal, "LXMERT: Learning Cross-Modality Encoder Representationsfrom Transformers", EMNLP 2019

Visual Commonsense Reasoning Leaderboard



hide all | show all | [person1] | [person2] | [person3] | [person4]

more objects »

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington* (Zellers et al. '18)	91.0	93.0	85.0
September 30, 2019	UNITER-large (ensemble) *MS D365 AI* https://arxiv.org/abs/1909.11740	**79.8**	**83.4**	**66.8**
2 September 23, 2019	UNITER-large (single model) *MS D365 AI* https://arxiv.org/abs/1909.11740	77.3	80.8	62.8
3 August 9,2019	ViLBERT (ensemble of 10 models) *Georgia Tech & Facebook AI Research* https://arxiv.org/abs/1908.02265	76.4	78.0	59.8
4 September 23,2019	VL-BERT (single model) *MSRA & USTC* https://arxiv.org/abs/1908.08530	75.8	78.4	59.7
5 August 9,2019	ViLBERT (ensemble of 5 models) *Georgia Tech & Facebook AI Research* https://arxiv.org/abs/1908.02265	75.7	77.5	58.8

https://visualcommonsense.com/leaderboard/

What can we learn from reconstructing the input?

Stanford University is located in ______________, California.

What can we learn from reconstructing the input?

I put ____ fork down on the table.

What can we learn from reconstructing the input?

The woman walked across the street,
checking for traffic over ____shoulder.

What can we learn from reconstructing the input?

I went to the ocean to see the fish, turtles, seals, and ______.

What can we learn from reconstructing the input?

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.

What can we learn from reconstructing the input?

Iroh went into the kitchen to make some tea.

Standing next to Iroh, Zuko pondered his destiny.

Zuko left the_______.

What can we learn from reconstructing the input?

I was thinking about the sequence that goes

1, 1, 2, 3, 5, 8, 13, 21, _____

Interlude:
What kinds of things does pretraining learn?

There's increasing evidence that pretrained models learn a wide variety of things about the statistical properties of language:

- *Stanford University is located in___________, California.* [trivia]
- *I put____fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over____shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and______.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was____.* [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the_______. [some reasoning – this is harder]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21,_____[some basic arithmetic; they don't learn the Fibonnaci sequence]
- Models also learn – and can exacerbate racism, sexism, all manner of bad biases.
- More on all this in the interpretability lecture!

Adapted from John Hewitt

Plan for this lecture

- Learning the relation between images and text
  - Recurrent neural networks
  - Applications: Captioning
  - Transformers
- Reasoning: Visual question answering
  - Neuro-symbolic VQA
  - Graph convolutional networks
- Multimodal self-supervised learning

Visual Question Answering and Visual Reasoning

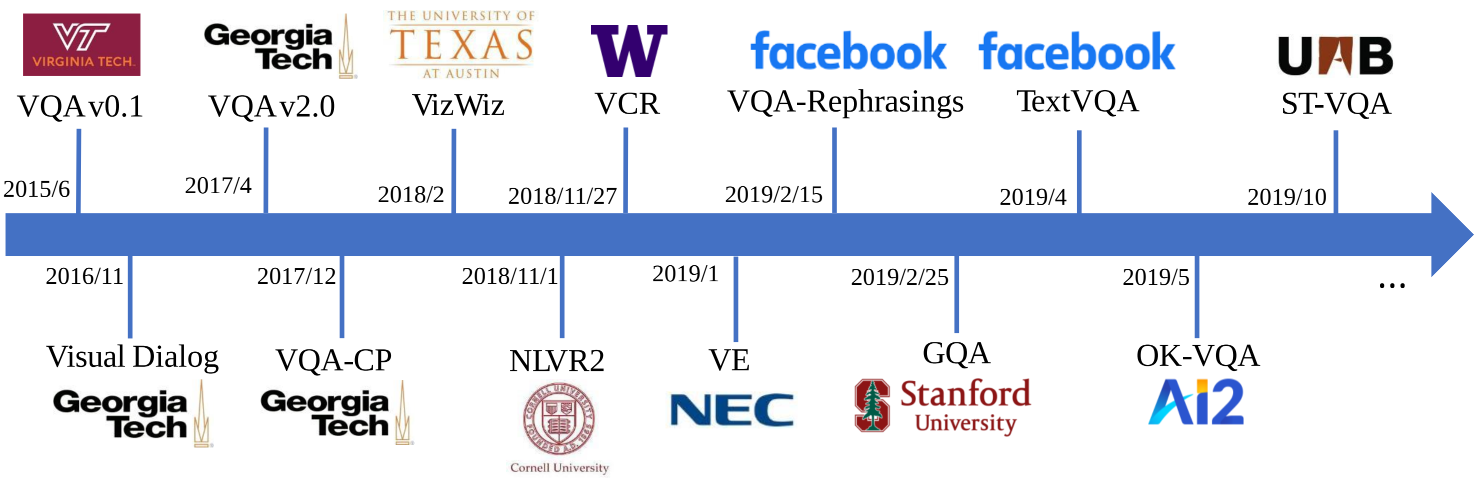
Zhe Gan

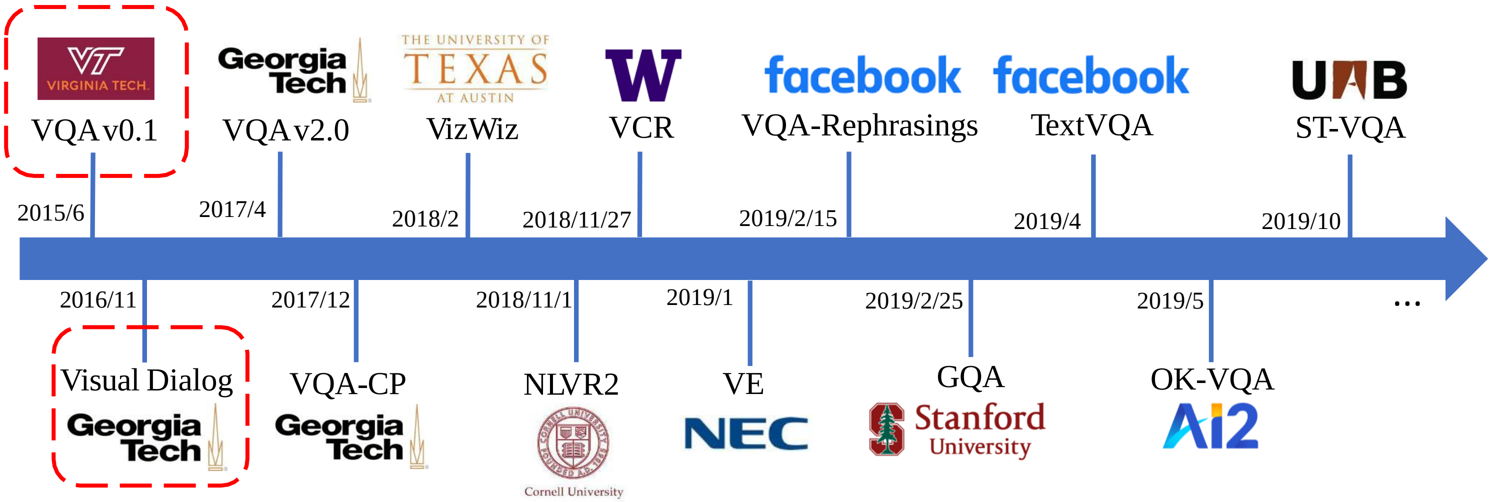
6/15/2020

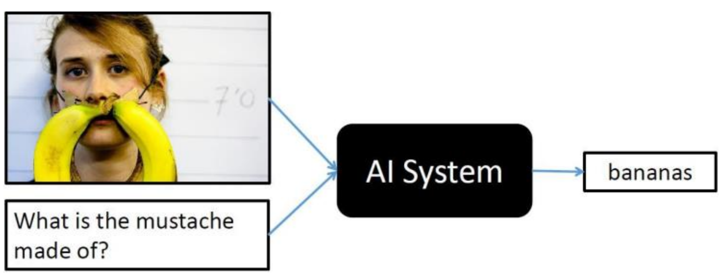
Microsoft

Task Overview: VQA and Visual Reasoning

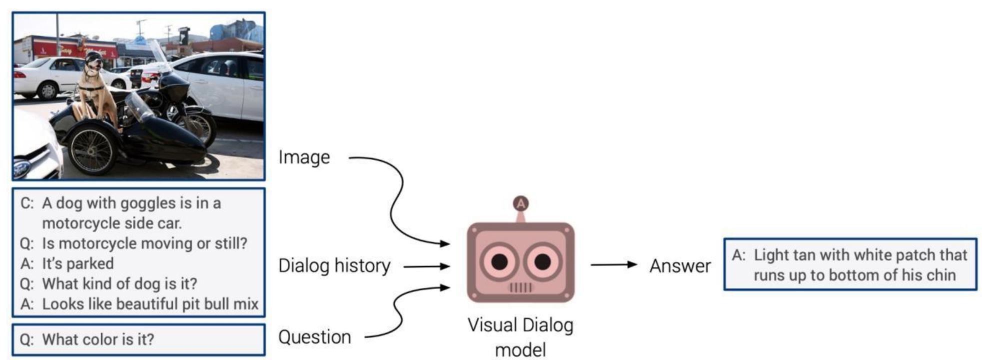
- Large-scale annotated datasets have driven tremendous progress in this field







VQA

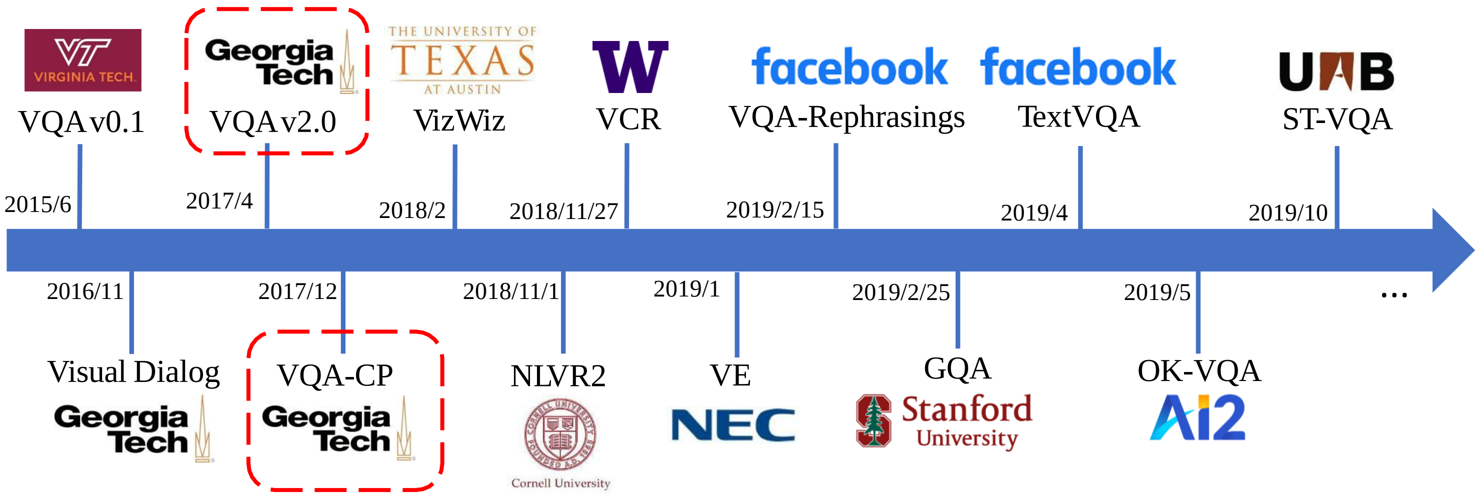


Visual Dialog

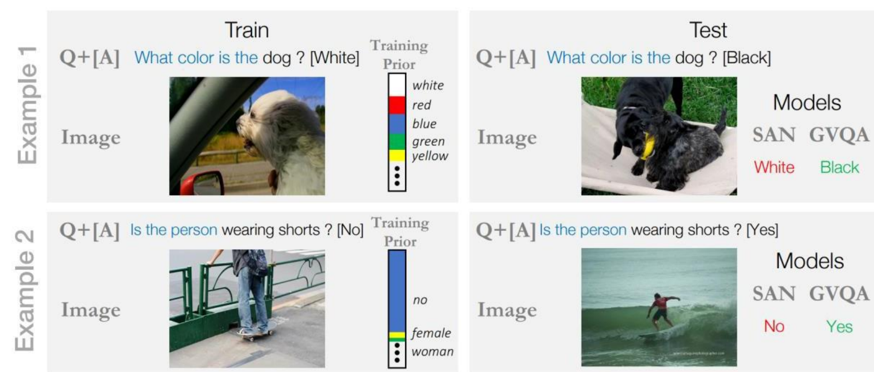
Image credit: https://visualqa.org/, https://visualdialog.org/

1 VQA: Visual Question Answering, ICCV 2015
2 Visual Dialog, CVPR 2017

Zhe Gan



VQA v2.0

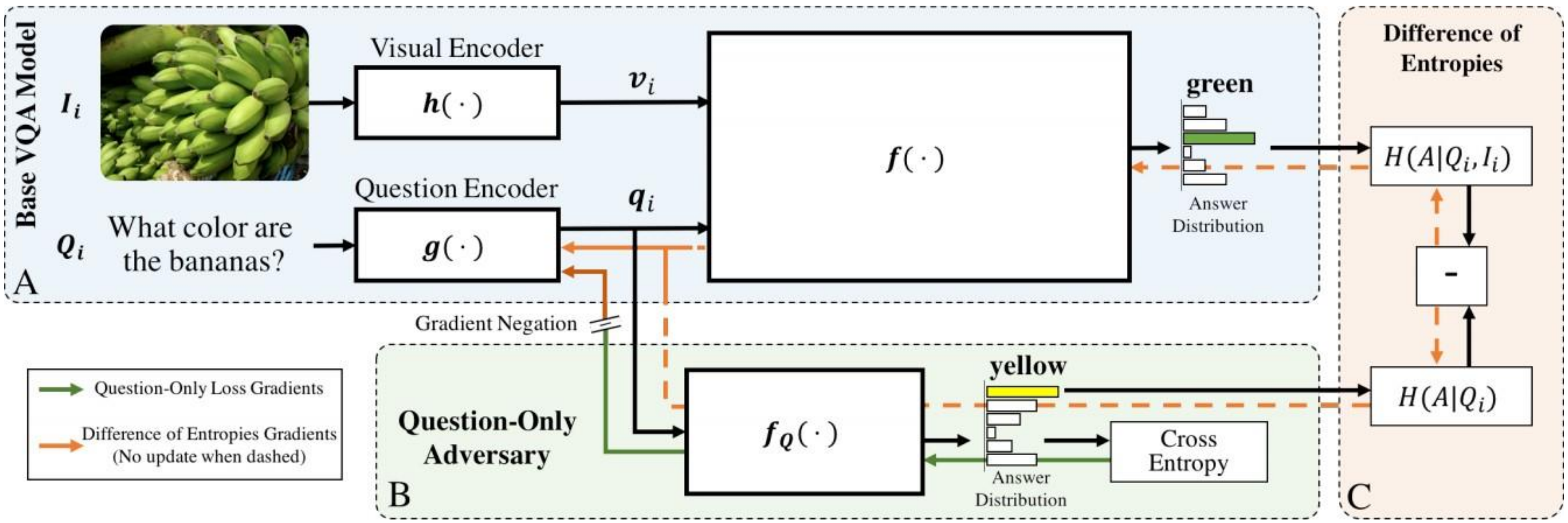


VQA-CP

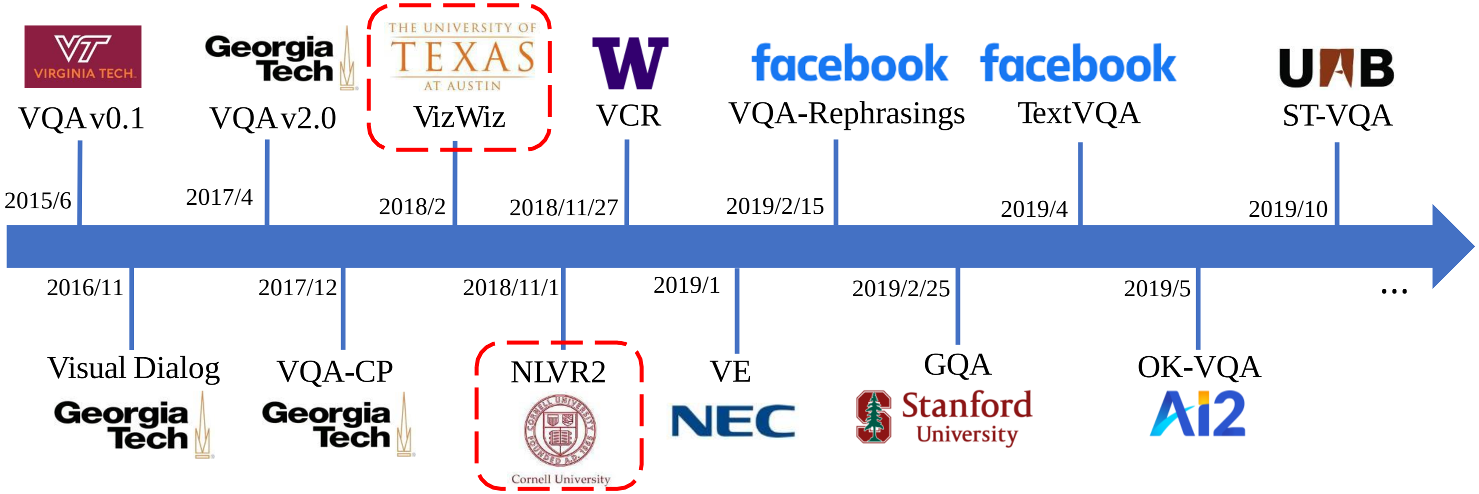
1 Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CVPR 2017
2 Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering, CVPR 2018

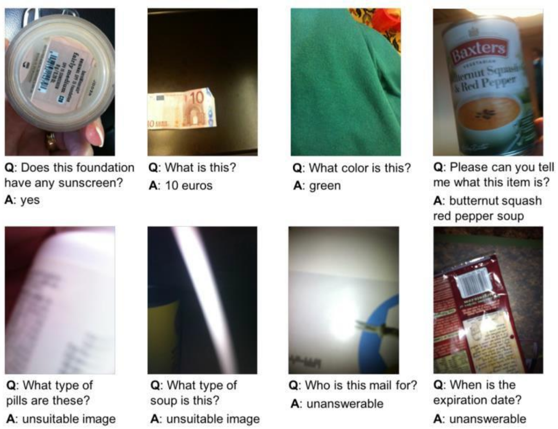
Robust VQA

- Overcoming language prior with adversarial regularization



[1] Overcoming Language Priors in Visual Question Answering with Adversarial Regularization, NeurIPS 2018

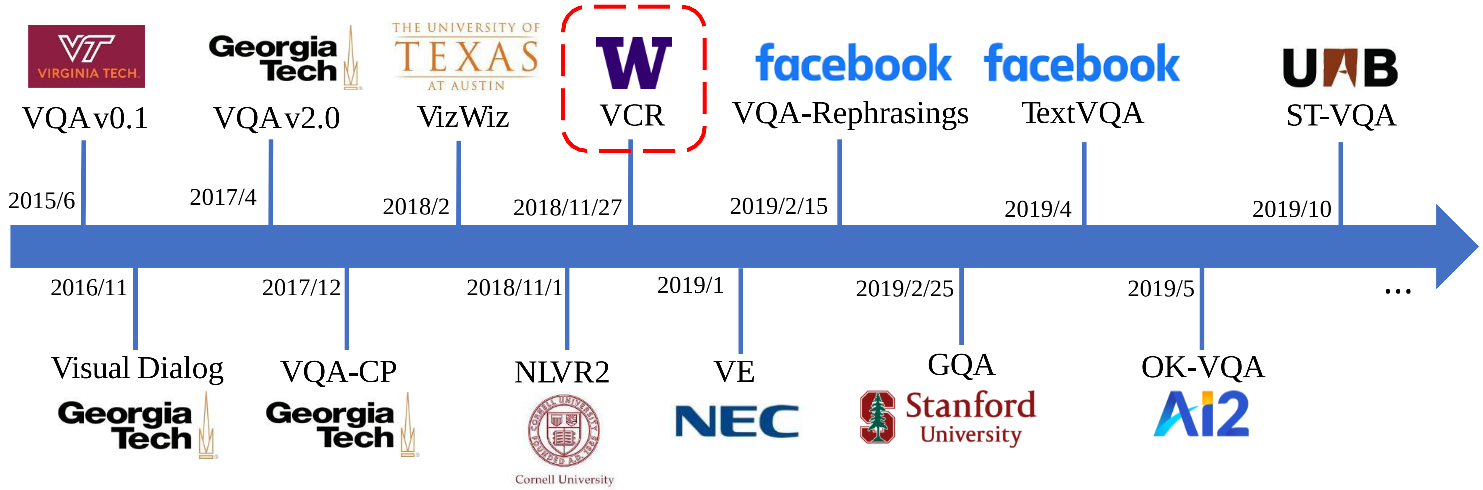


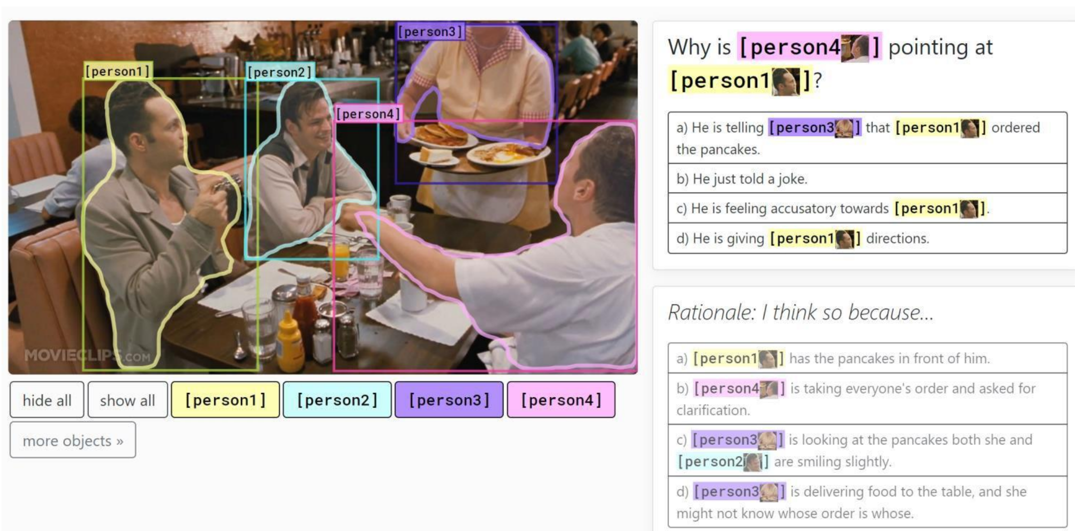


VizWiz

NLVR2

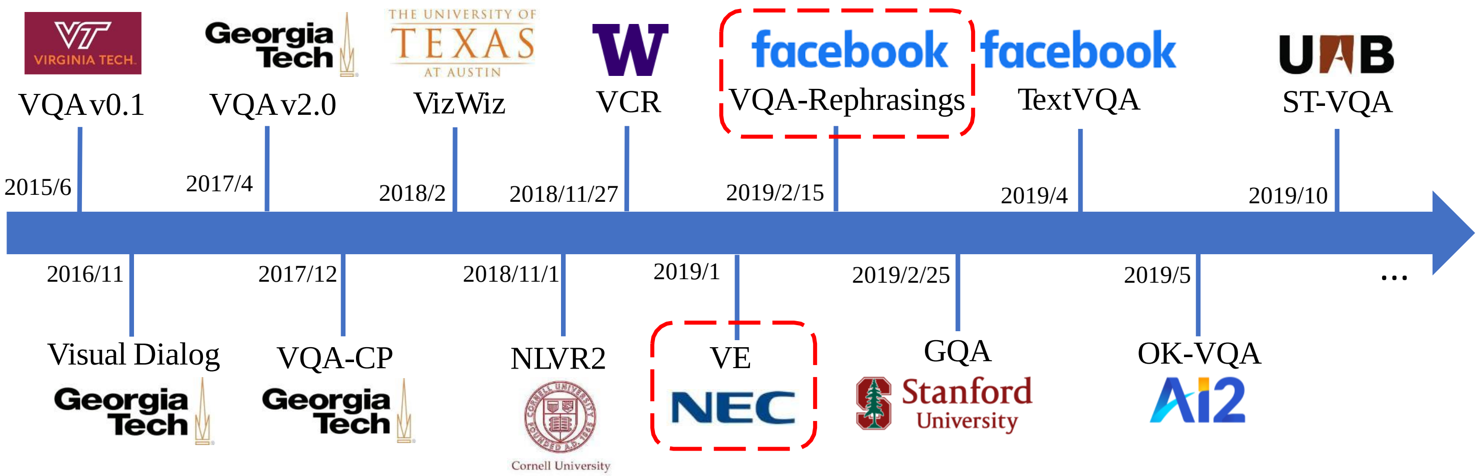
1 VizWiz Grand Challenge: Answering Visual Questions from Blind People, CVPR 2018
2 A Corpus for Reasoning About Natural Language Grounded in Photographs, ACL 2019

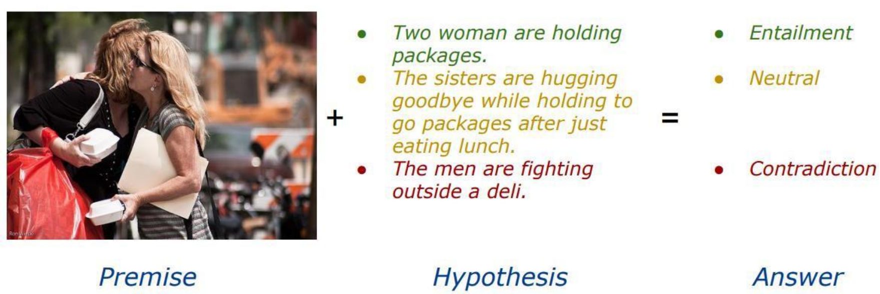




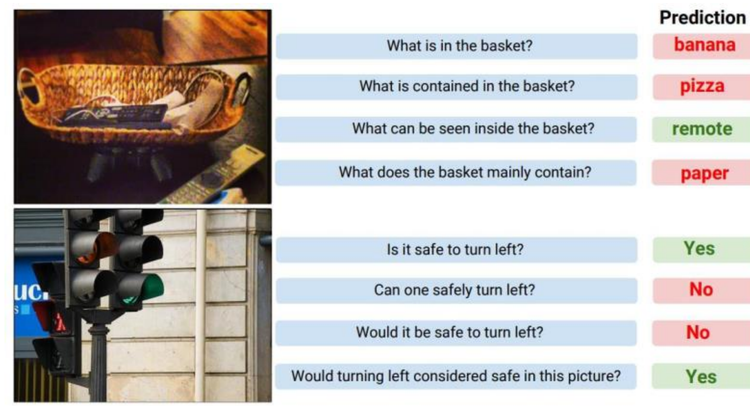
[1] From Recognition to Cognition: Visual Commonsense Reasoning, CVPR 2019

Zhe Gan



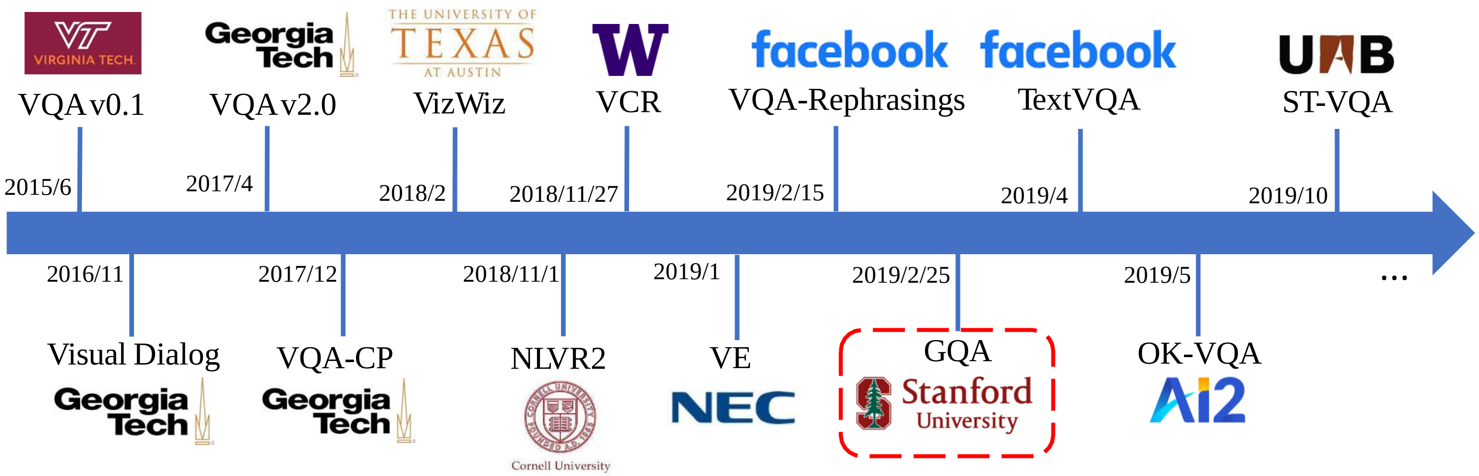


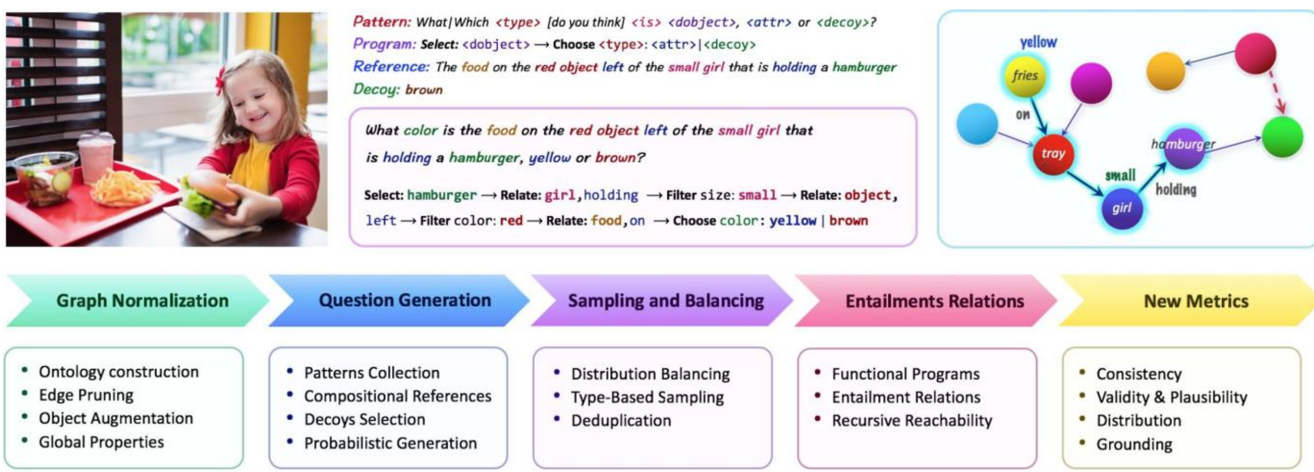
Visual Entailment



VQA-Rephrasings

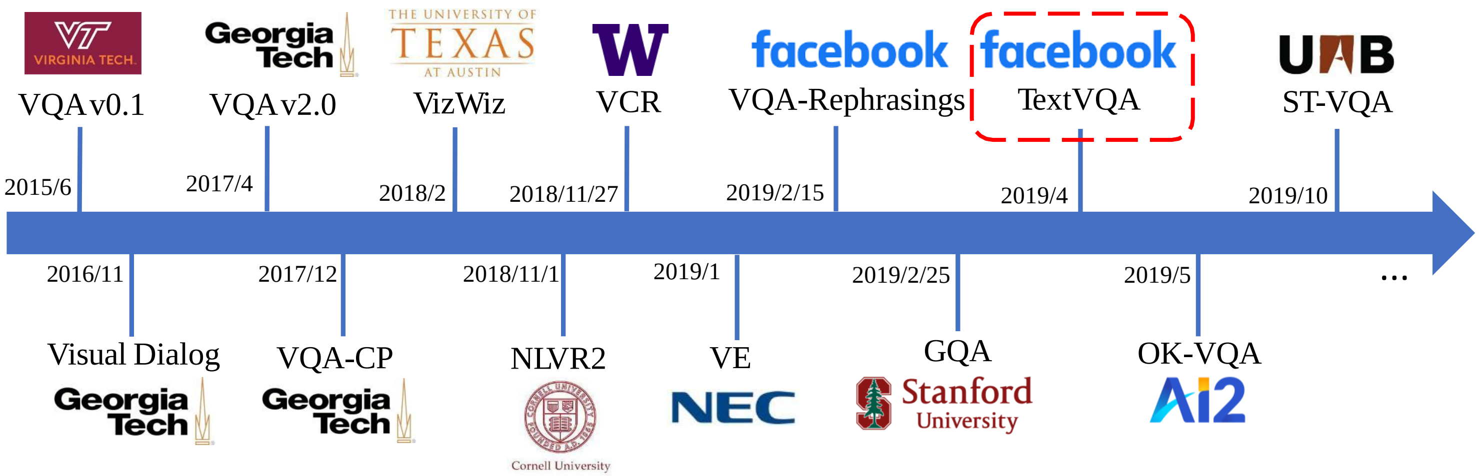
1 Visual Entailment: A Novel Task for Fine-Grained Image Understanding, 2019
2 Cycle-Consistency for Robust Visual Question Answering, CVPR2019





[1] GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, CVPR2019

Zhe Gan



What is the top oz?

Ground Truth	Prediction
16	red

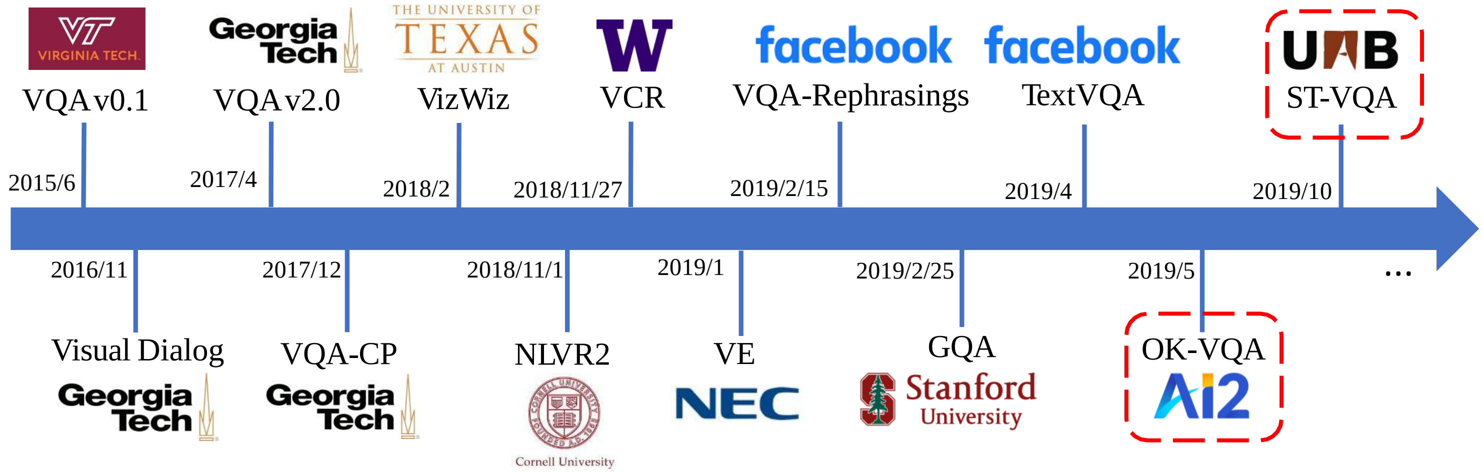
What is the largest denomination on table?

Ground Truth	Prediction
500	unknown

TextVQA

A dataset to benchmark visual reasoning based on text in images.

[1] Towards VQA Models That Can Read, CVPR2019



Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

OK-VQA

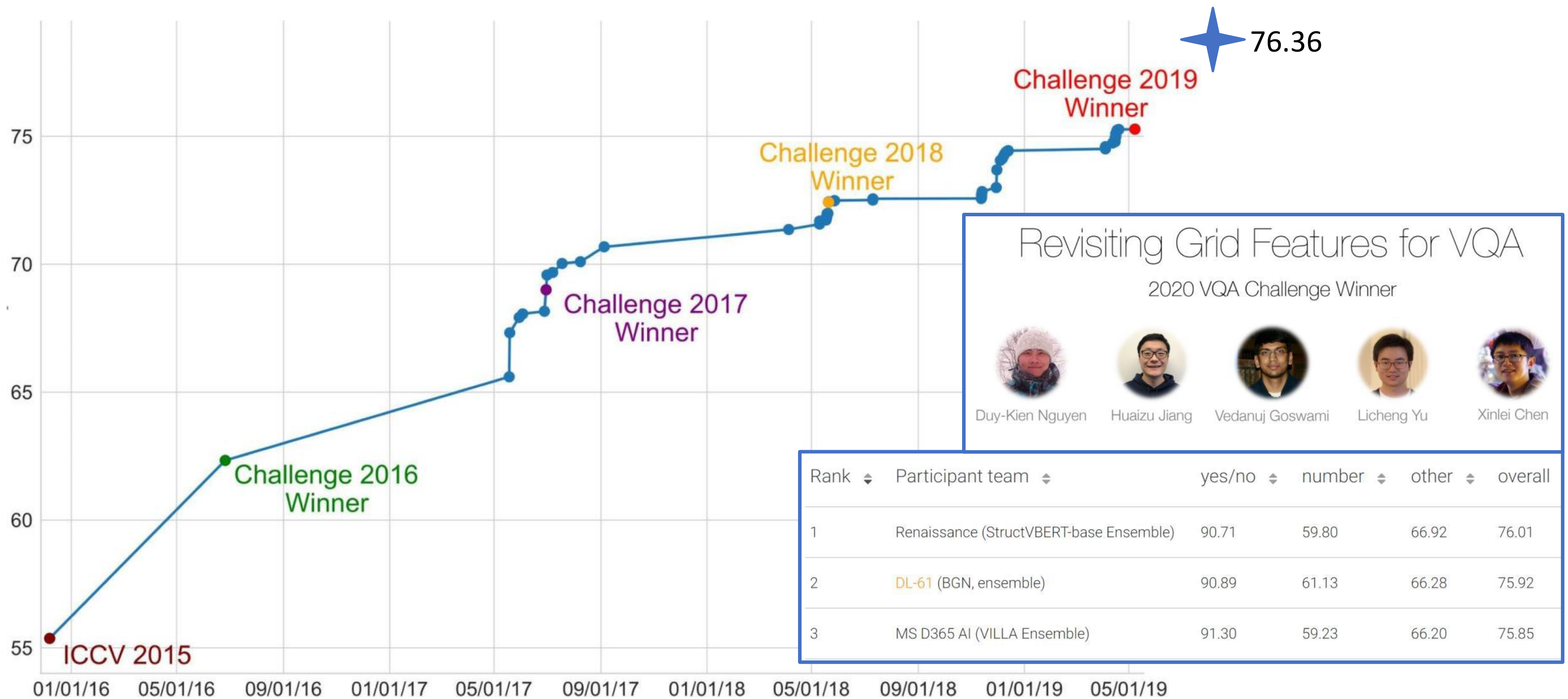
Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

Scene Text VQA

1 OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, CVPR 2019
2 Scene Text Visual Question Answering, ICCV 2019

Visual Question Answering



Rank	Participant team	yes/no	number	other	overall
1	Renaissance (StructVBERT-base Ensemble)	90.71	59.80	66.92	76.01
2	DL-61 (BGN, ensemble)	90.89	61.13	66.28	75.92
3	MS D365 AI (VILLA Ensemble)	91.30	59.23	66.20	75.85

Image Credit: CVPR 2019 Visual Question Answering and Dialog Workshop

More datasets…

SQuINTing at VQA Models: Interrogating VQA Models with Sub-Questions

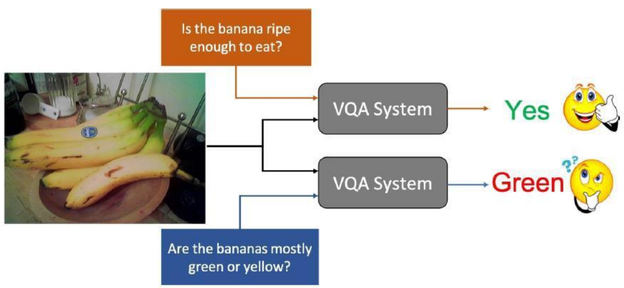
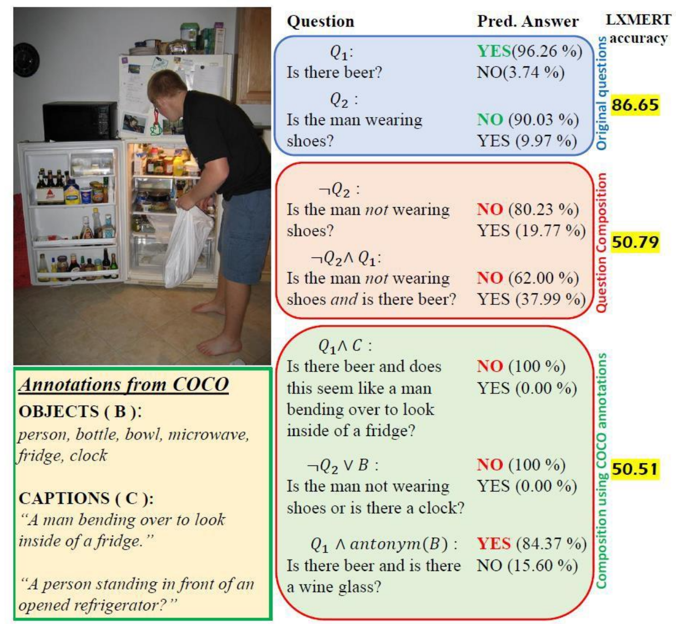


Figure 1: A potential reasoning failure: Current models answer "Yes" correctly to the Reasoning question "Is the banana ripe enough to eat?". We might assume that correctly answering the Reasoning question stems from perceiving relevant concepts correctly – perceiving yellow bananas in this example. But when asked "Are the bananas mostly green or yellow?", it answers "Green" incorrectly – indicating that the model possibly answered the original for the wrong reasons even if the answer was right. We quantify the extent to which this phenomenon occurs in VQA and introduce a new dataset aimed at stimulating research on well grounded reasoning.

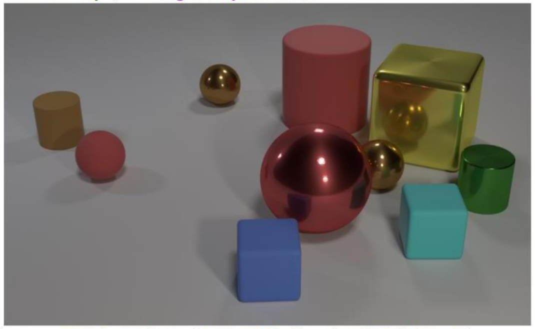
VQA-LOL: Visual Question Answering under the Lens of Logic



Diagnostic Datasets

- CLEVR (Compositional Language and Elementary Visual Reasoning)
  - Has been extended to visual dialog (CLEVR-Dialog), referring expressions (CLEVR-Ref+), and video reasoning (CLEVRER)

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.

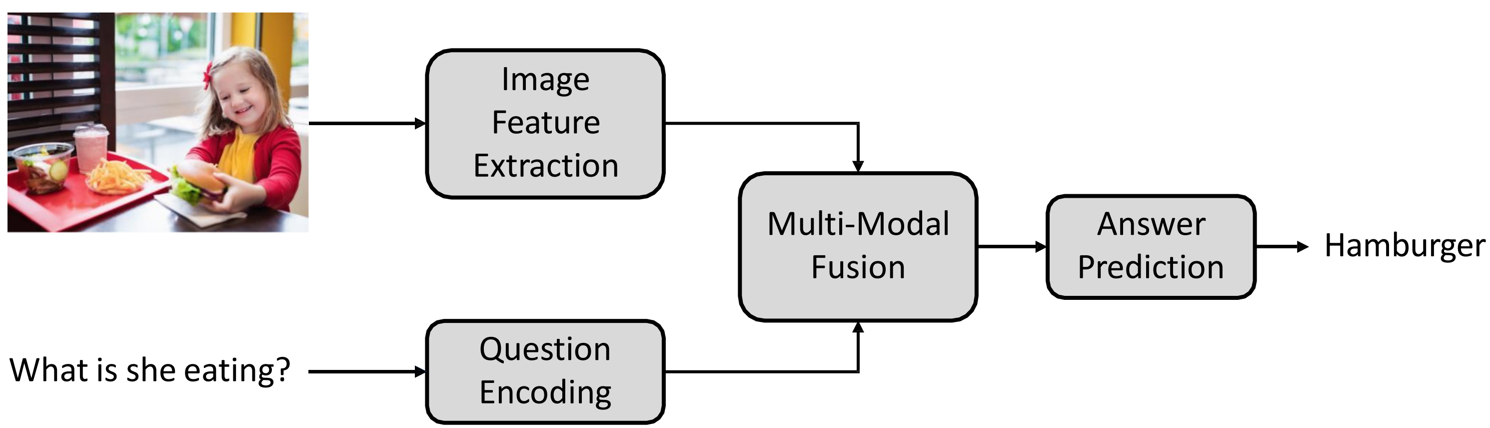


Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either small cylinders** or **red** things?

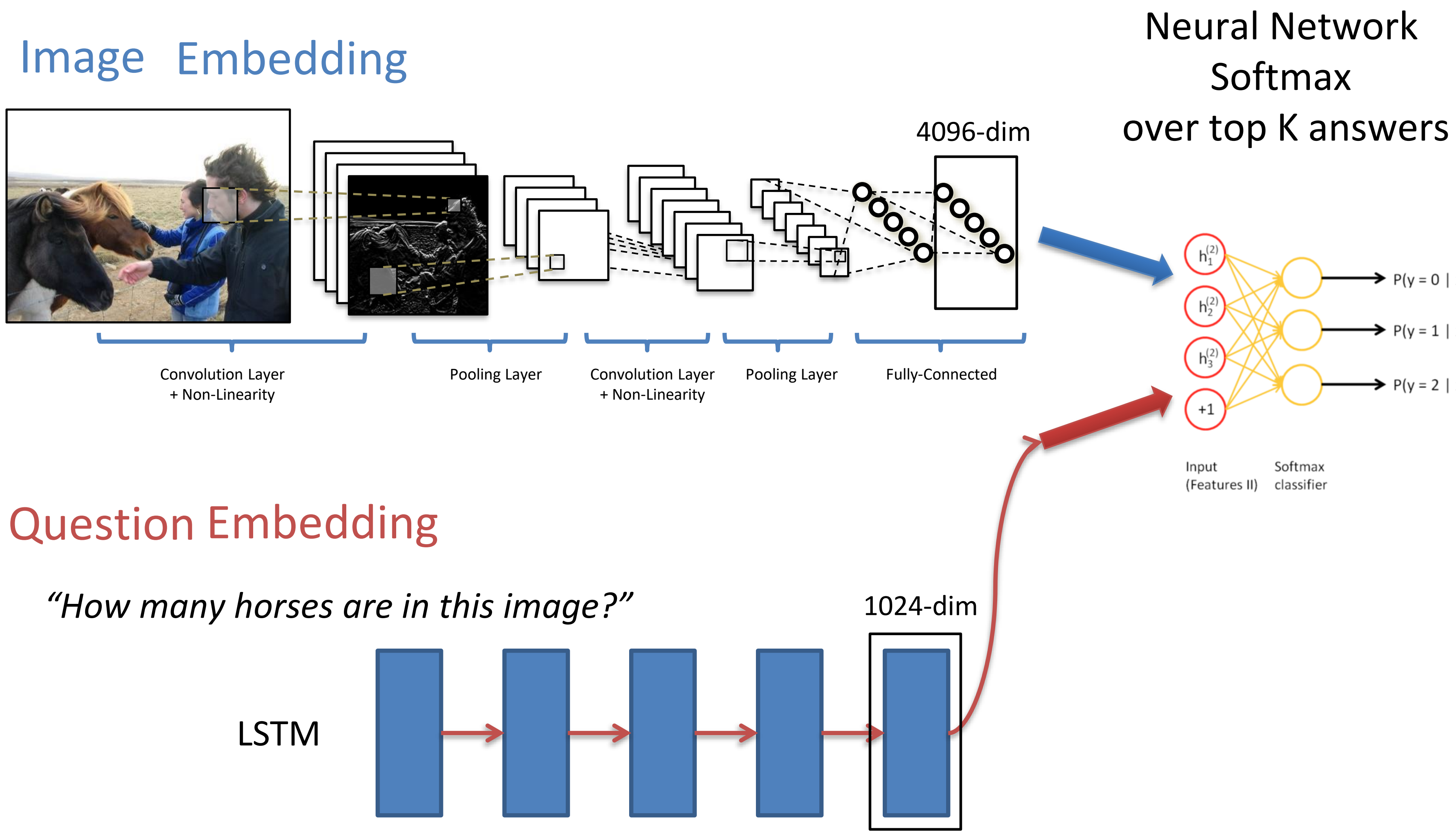
1 CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017
2 CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog, NAACL 2019
3 CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions, CVPR 2019
4 CLEVRER: CoLlision Events for Video REpresentation and Reasoning, ICLR 2020

Overview

- What a typical system looks like



Example VQA system



Agrawal et al., "[VQA: Visual Question Answering]", ICCV 2015

Example VQA system

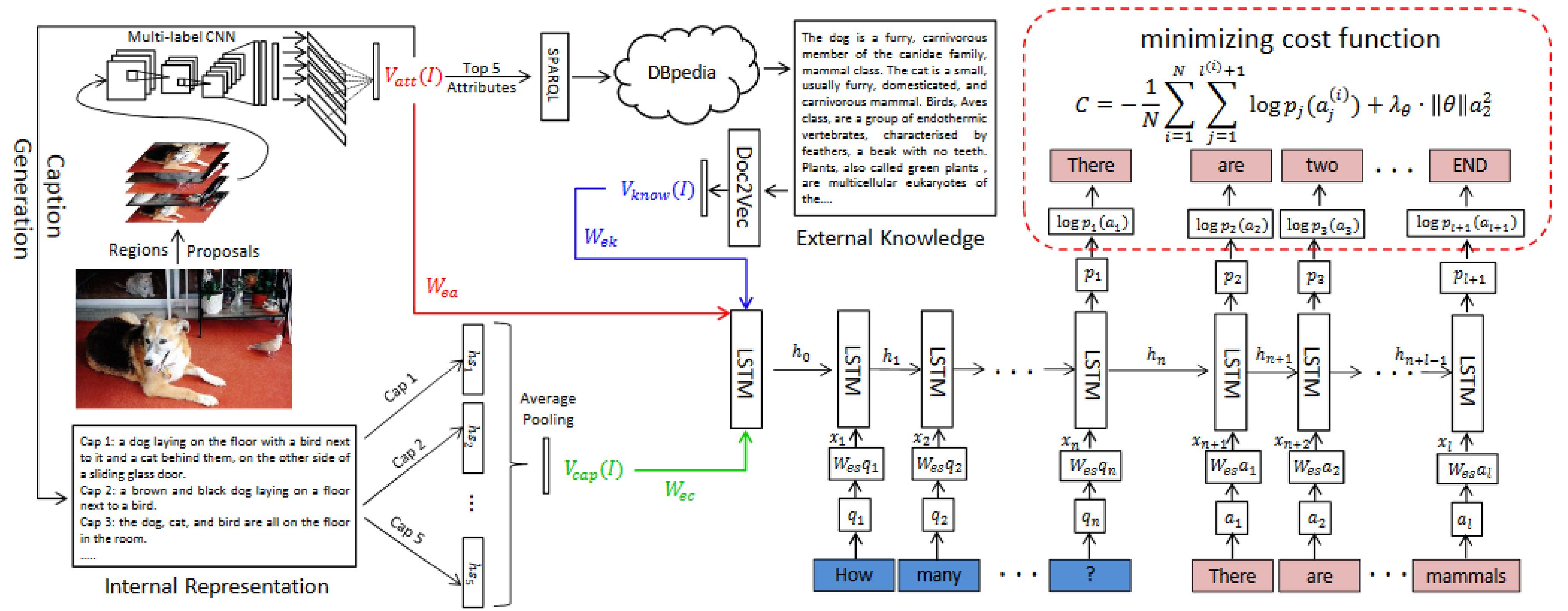


Figure 2. Our proposed framework: given an image, a CNN is first applied to produce the attribute-based representation $V_{att}(I)$. The internal textual representation is made up of image captions generated based on the image-attributes. The hidden state of the caption-LSTM after it has generated the last word in each caption is used as its vector representation. These vectors are then aggregated as $V_{cap}(I)$ with average-pooling. The external knowledge is mined from the KB (in this case DBpedia) and the responses encoded by Doc2Vec, which produces a vector $V_{know}(I)$. The 3 vectors $\mathbf{V}$ are combined into a single representation of scene content, which is input to the VQA LSTM model which interprets the question and generates an answer.

Wu et al., "Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge From External Sources", CVPR 2016

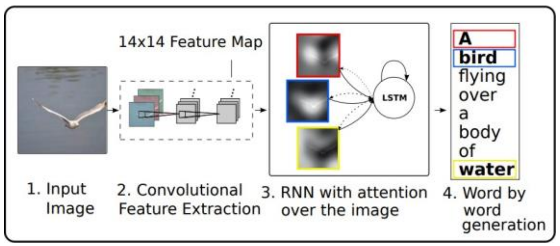
Mila — Show, Attend and Tell — 2015/2

Microsoft — SAN — 2015/11

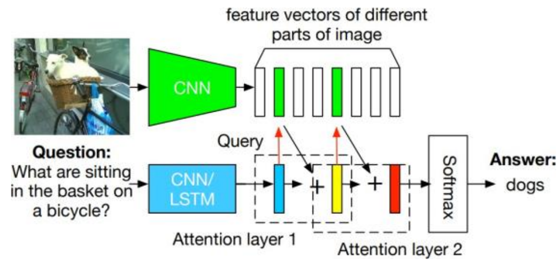
Microsoft — BUTD — 2017/7

facebook — Grid Feature — 2020/1

Microsoft — Pixel-BERT — 2020/4



Show, Attend and Tell



Stacked Attention Network

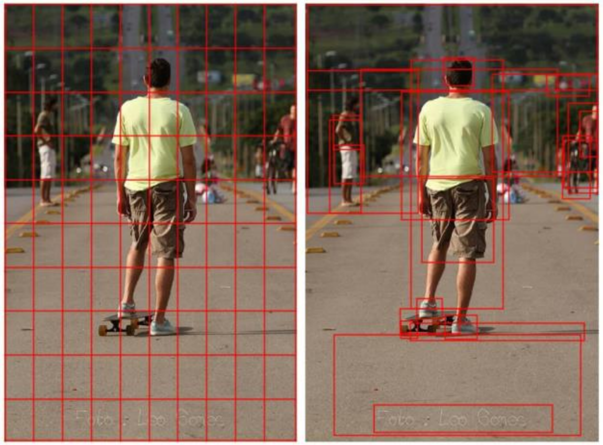
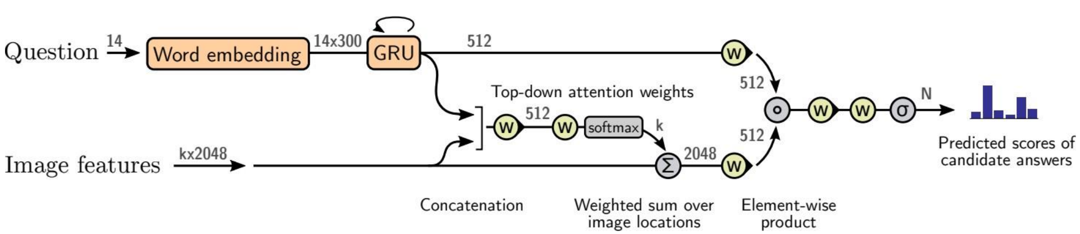


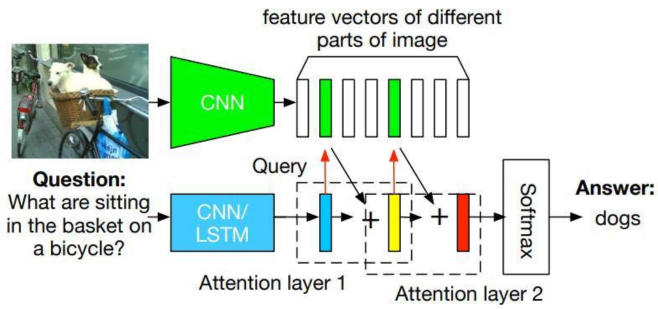
Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).



2017 VQA Challenge Winner

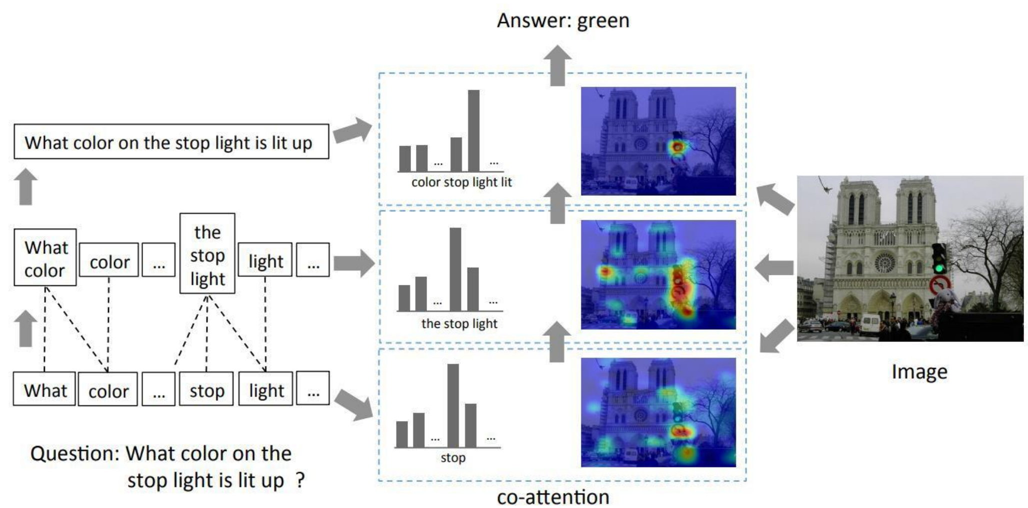
1 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015
2 Stacked Attention Networks for Image Question Answering, CVPR 2016
3 Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR 2018

Zhe Gan



(a) Stacked Attention Network for Image QA

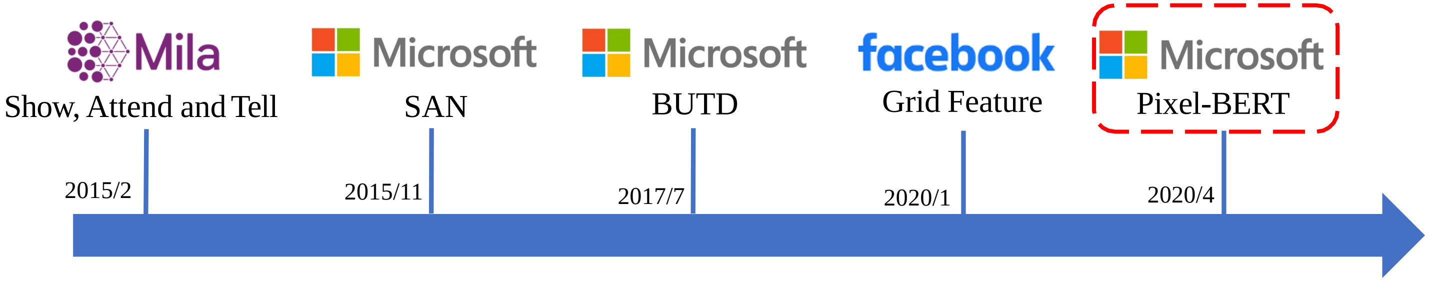
(b) Visualization of the learned multiple attention layers.

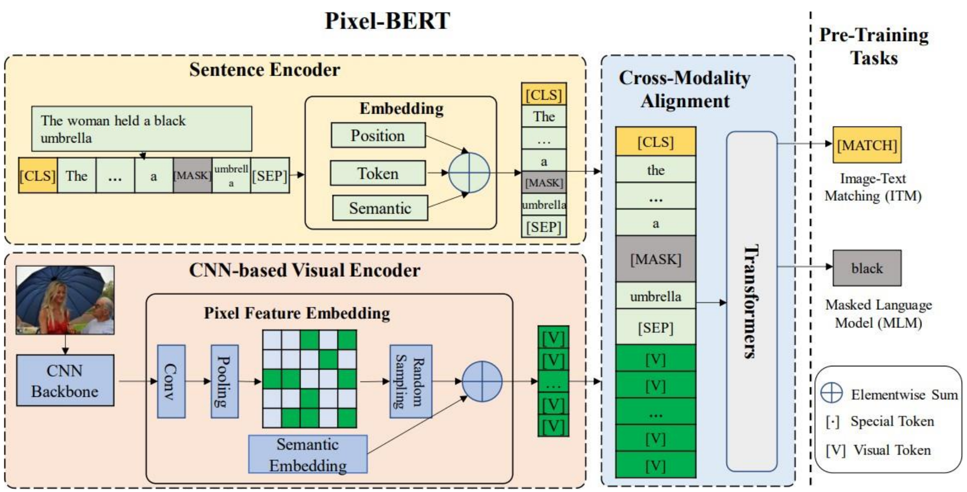


Parallel Co-attention and Alternative Co-attention

1 Stacked Attention Networks for Image Question Answering, CVPR 2016
2 Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016

Zhe Gan





Model	test-dev	test-std
MUTAN[5]	60.17	-
BUTD[2]	65.32	65.67
ViLBERT[21]	70.55	70.92
VisualBERT[19]	70.80	71.00
VLBERT[29]	71.79	72.22
LXMERT[33]	72.42	72.54
UNITER[6]	72.27	72.46
Pixel-BERT (r50)	71.35	71.42
Pixel-BERT (x152)	**74.45**	**74.55**

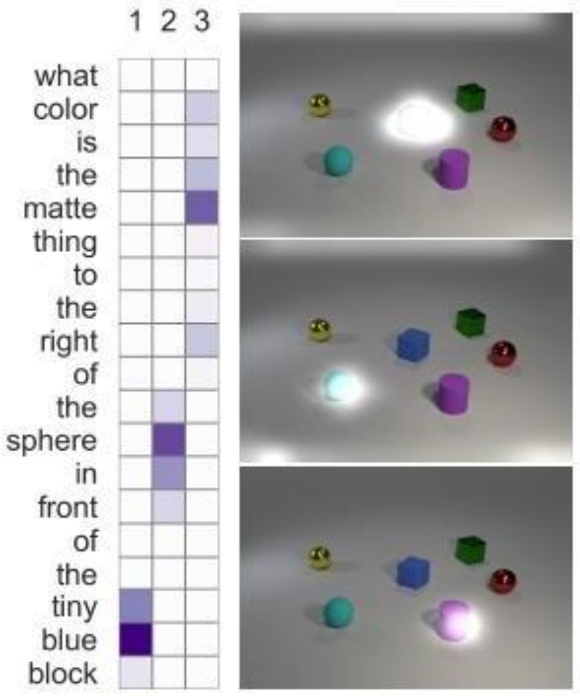
Table 2. Evaluation of Pixel-BERT with other methods on VQA.

[1] Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers,2020

Zhe Gan

MAC: Memory, Attention and Composition

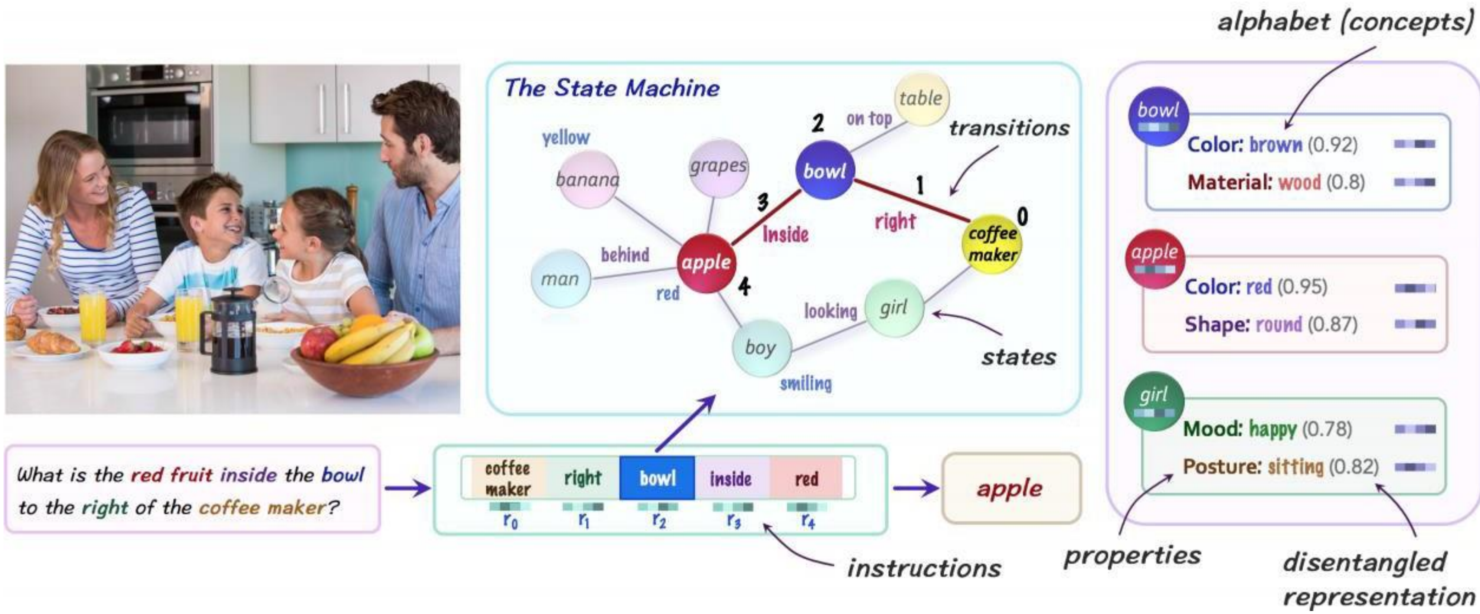
- Each cell maintains recurrent dual states:
  - *Control c_i*: the reasoning operation that should be accomplished at this step.
  - *Memory m_i*: the retrieved information relevant to the query, accumulated over previous iterations.
  - *Implementation-wise*:
    - **Attention-based average** of a given query (question)
    - **Attention-based average** of a given Knowledge Base (image)



[1] Compositional Attention Networks for Machine Reasoning, ICLR, 2018

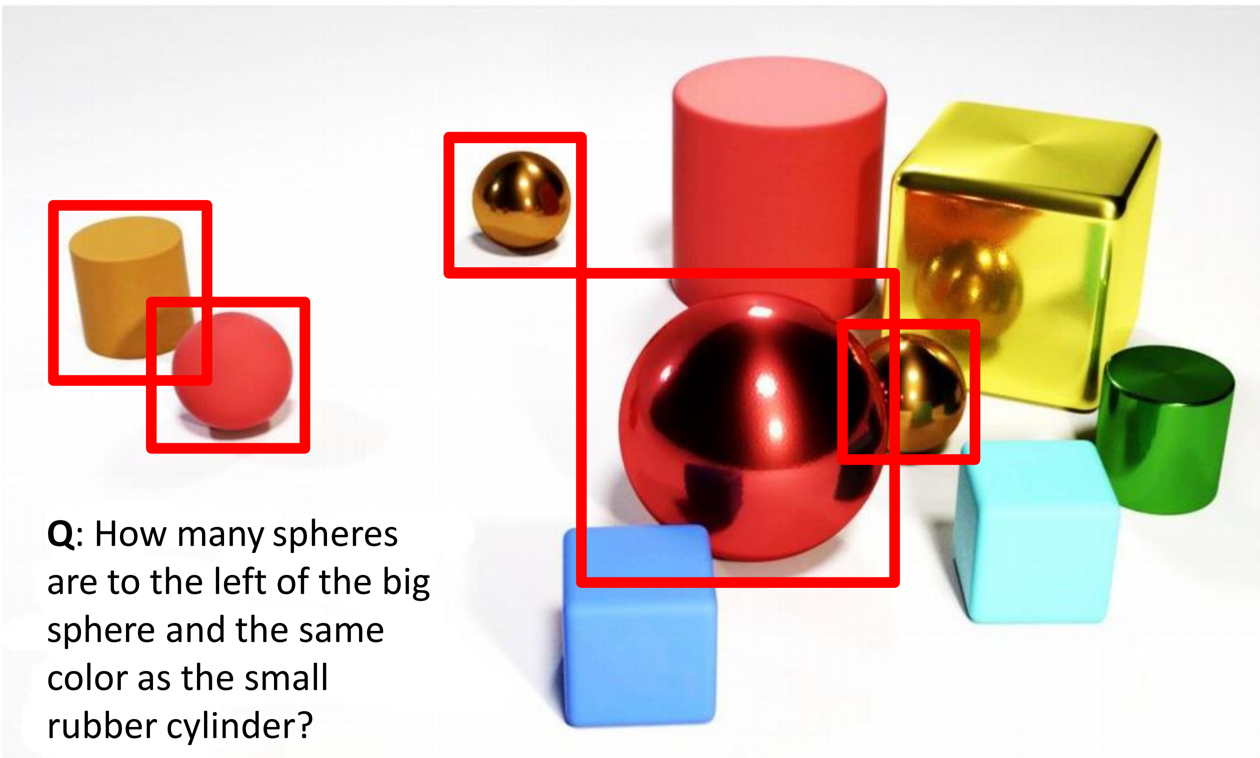
Neural State Machine

- We see and reason with concepts, not visual details, 99% of the time
- We build semantic world models to represent our environment



[1] Learning by Abstraction: The Neural State Machine, NeurIPS 2019

Compositional Visual Reasoning



Q: How many spheres are to the left of the big sphere and the same color as the small rubber cylinder?

Identify big sphere

↓

Spheres on left

↓

Rubber cylinder

↓

Sphere of same color

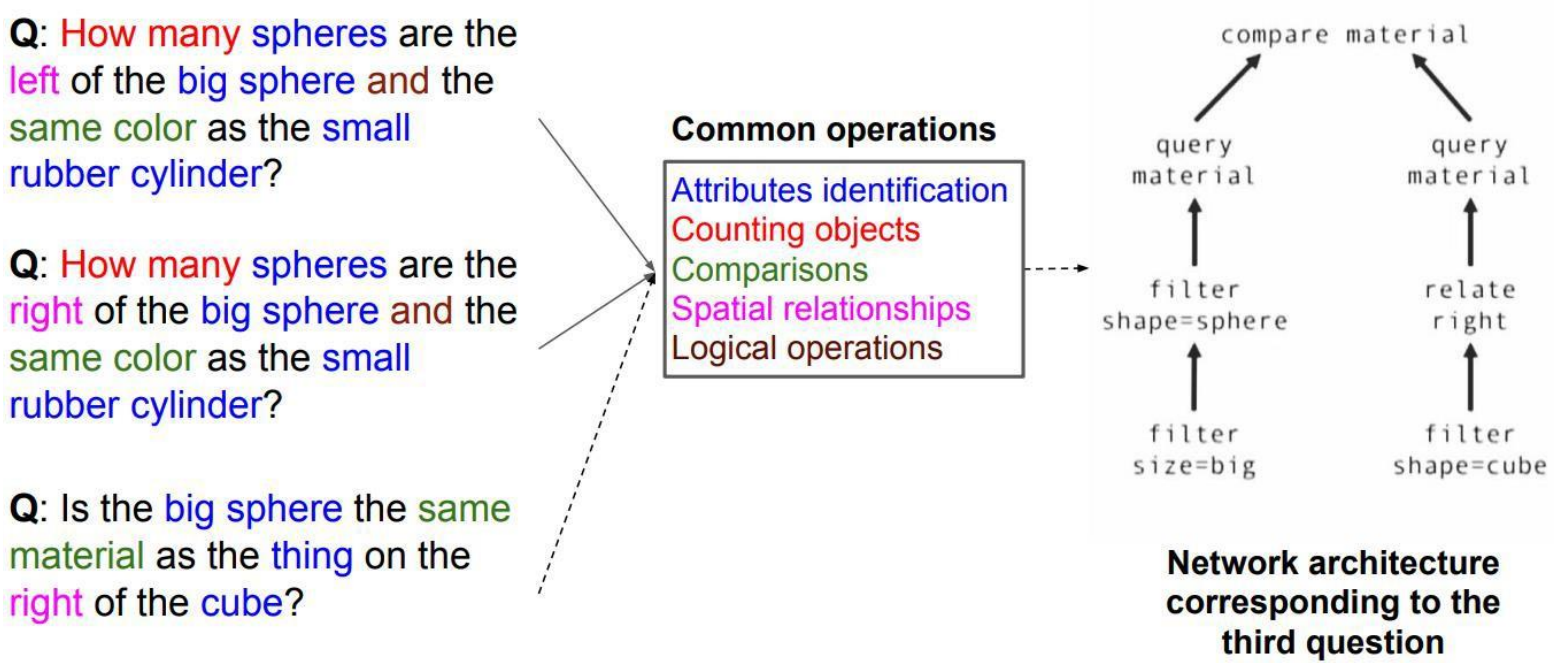
↓

Count

A: 1

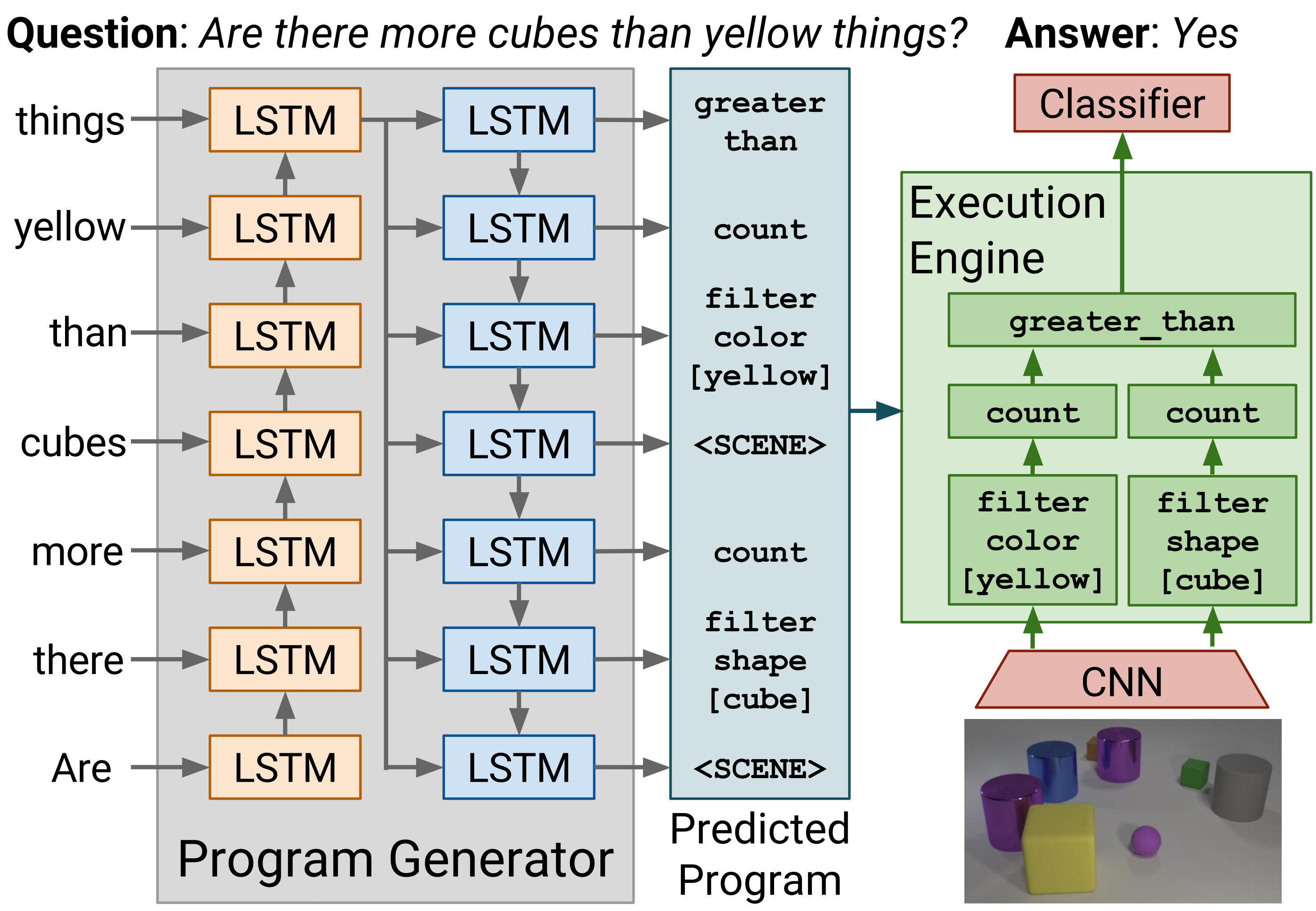
[1] CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, CVPR,2017

Consider a compositional model



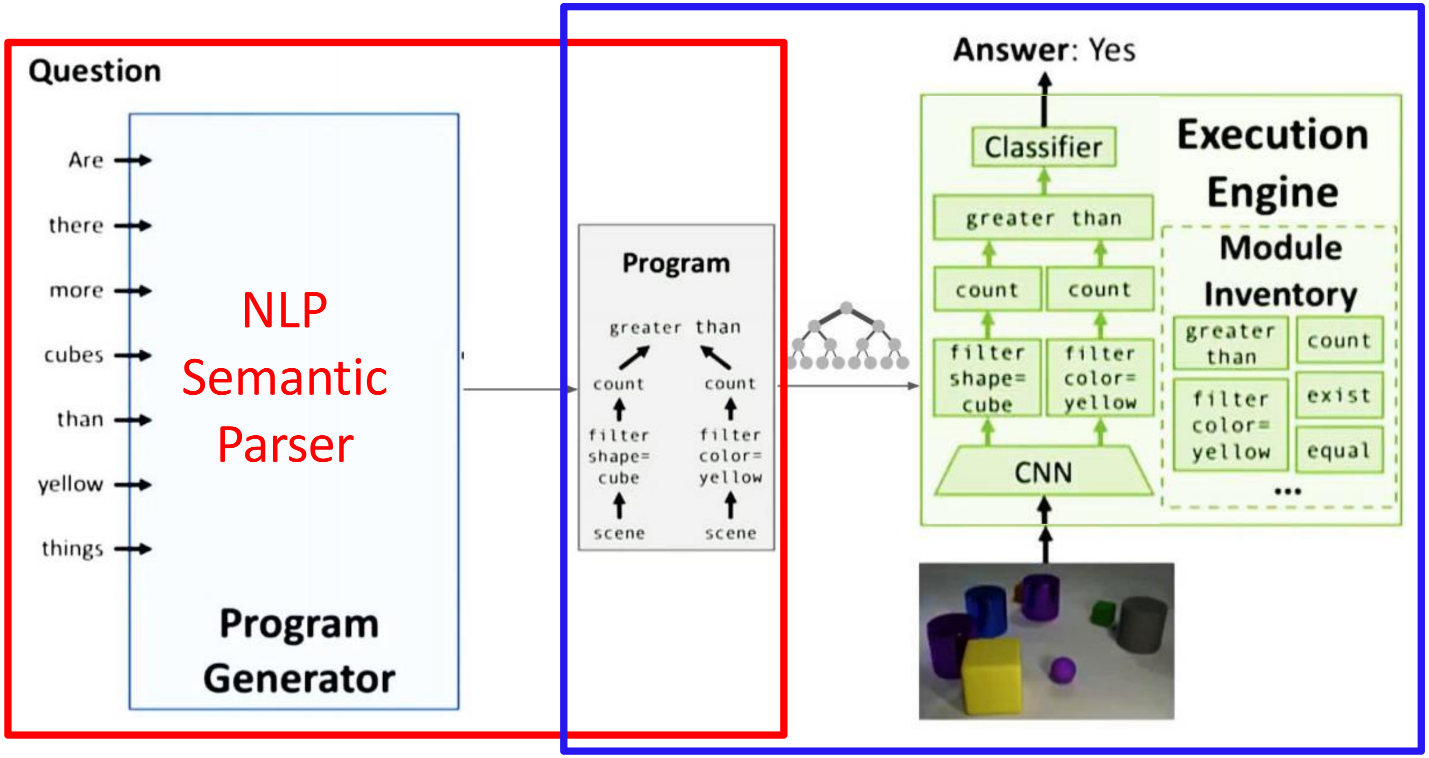
[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Overview of one compositional model



Johnson et al., "Inferring and Executing Programs for Visual Reasoning", ICCV 2017

Overview of one compositional model

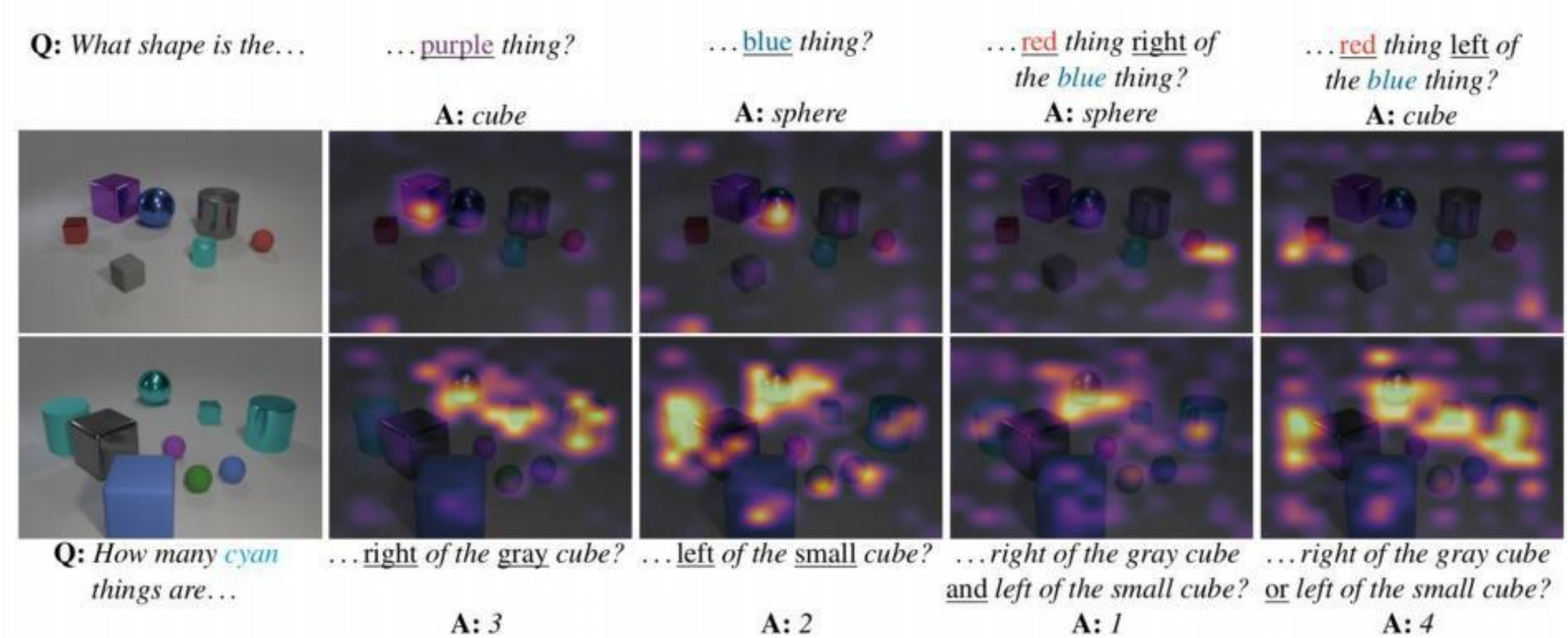


Uses some pre-trained parser

Trained separately

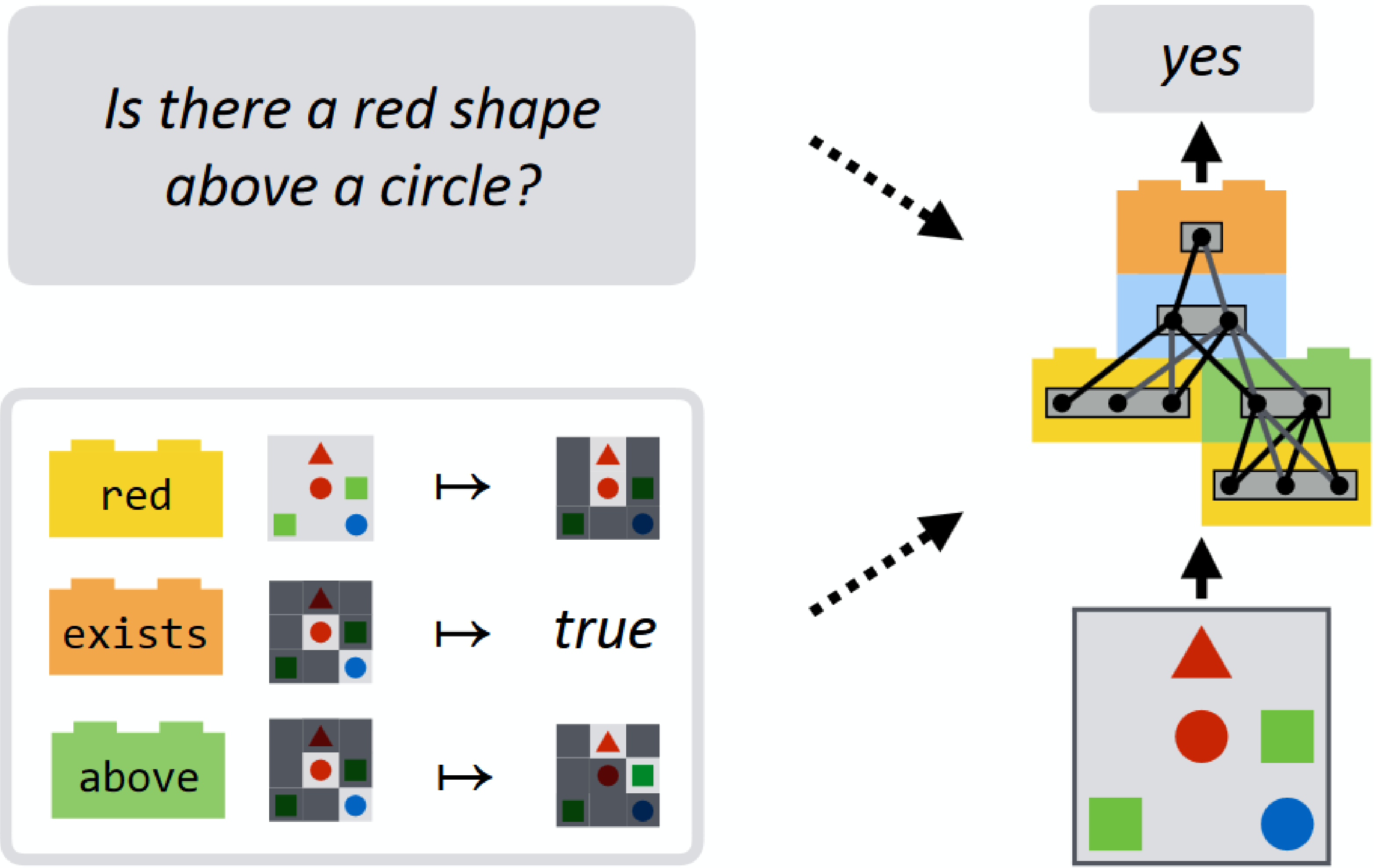
Johnson et al., "Inferring and Executing Programs for Visual Reasoning", ICCV 2017

What do the modules learn?



Johnson et al., "[Inferring and Executing Programs for Visual Reasoning]", ICCV 2017

Another compositional model

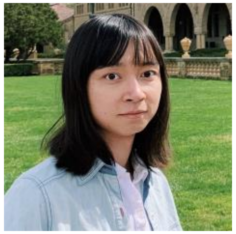


Andreas et al., "Neural Module Networks", CVPR 2016

The Neuro-Symbolic Concept Learner

Interpreting Scenes, Words, and Sentences From Natural Supervision

http://nscl.csail.mit.edu

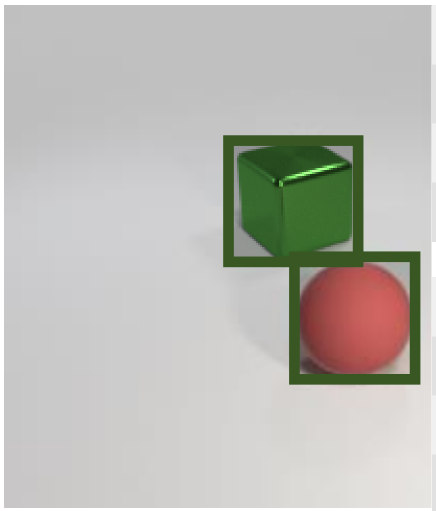




Jiayuan Mao[1,2] Chuang Gan[3] Pushmeet Kohli[4] Josh Tenenbaum[1] Jiajun Wu[1]

[1]MIT CSAIL [2]Tsinghua University [3]MIT-IBM Watson AI Lab [4]DeepMind

Concepts in Visual Reasoning

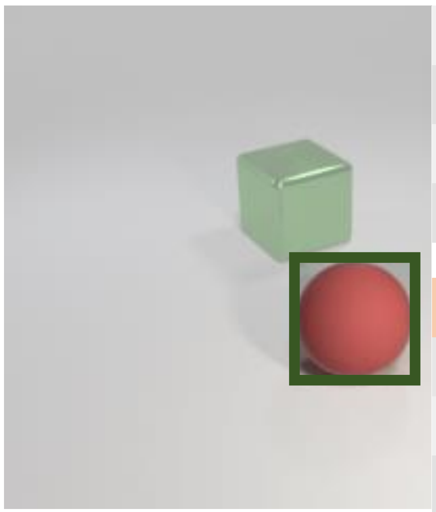


Color	Green
Shape	Cube
Material	Metal
……	……

Color	Red
Shape	Sphere
Material	Rubber
……	……

CLEVR [Johnson et al., 2017]

Concepts in Visual Reasoning



Color	Green
Shape	Cube
Material	Metal
……	……

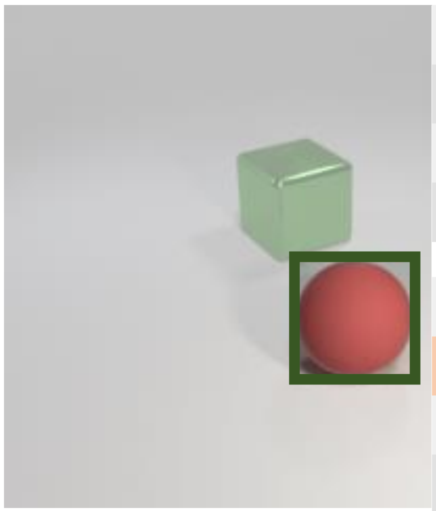
Color	Red
Shape	Sphere
Material	Rubber
……	……

CLEVR [Johnson et al., 2017]

Visual Question Answering

Q: What's the shape of the **red** object?

Concepts in Visual Reasoning



Color	Green
Shape	Cube
Material	Metal
……	……

Color	Red
Shape	Sphere
Material	Rubber
……	……

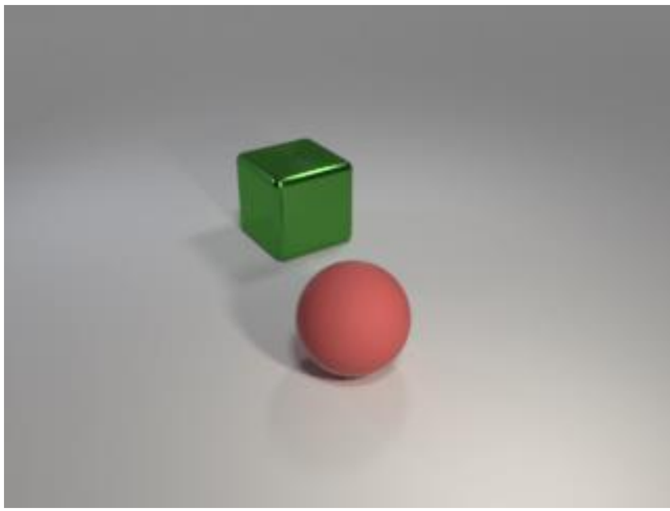
Visual Question Answering

Q: What’s the **shape** of the **red** object?

A: Sphere.

CLEVR [Johnson et al., 2017]

End-to-End Visual Reasoning



Visual Question Answering

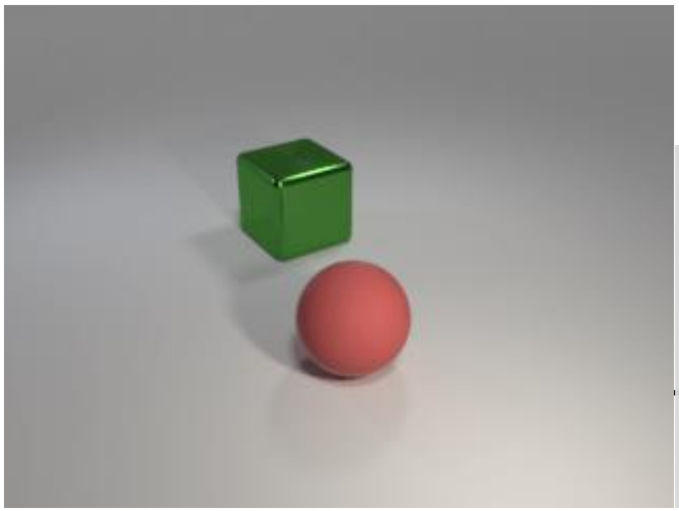
Q: What's the **shape** of the **red** object?

End-to-End
Neural Network

A: Sphere.

NMN [Andreas et al., 2016]
IEP [Johnson et al., 2017]
FiLM [Perez et al., 2018],
MAC [Hudson & Manning, 2018]
Stack-NMN [Hu et al., 2018]
TbD [Mascharka et al. 2018]

End-to-End Visual Reasoning



Visual Question Answering
Q: What's the **shape** of the **red** object?

End-to-End
Neural Network

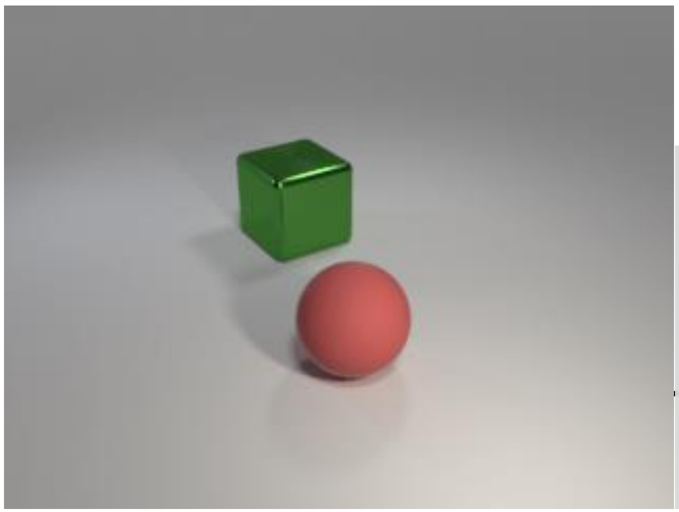
A: Sphere.

Concept
(e.g., colors, shapes)

Reasoning
(e.g., count)

NMN [Andreas et al., 2016]
IEP [Johnson et al., 2017]
FiLM [Perez et al., 2018],
MAC [Hudson & Manning, 2018]
Stack-NMN [Hu et al., 2018]
TbD [Mascharka et al. 2018]

End-to-End Visual Reasoning



Visual Question Answering

Q: What's the **shape** of the **red** object?

Entangled

Concept ⟷ Reasoning

(e.g., colors, shapes) (e.g., count)

Hard to transfer

Image Captioning

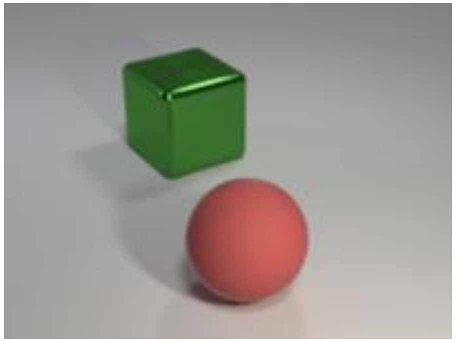
Instance Retrieval

NMN [Andreas et al., 2016]
IEP [Johnson et al., 2017]
FiLM [Perez et al., 2018],
MAC [Hudson & Manning, 2018]
Stack-NMN [Hu et al., 2018]
TbD [Mascharka et al. 2018]

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision



Scene Parsing →

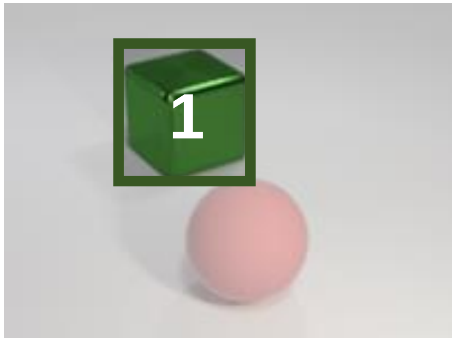
Language

Q: What's the **shape** of the **red** object?

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision



Scene Parsing

ID	Color	Shape	Material
1	Green	Cube	Metal

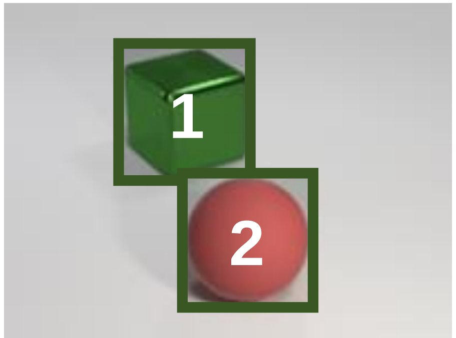
Language

Q: What's the **shape** of the **red** object?

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision



Scene Parsing →

ID	Color	Shape	Material
1	Green	Cube	Metal
2	Red	Sphere	Rubber

Language

Q: What's the **shape** of the **red** object?

Semantic Parsing →

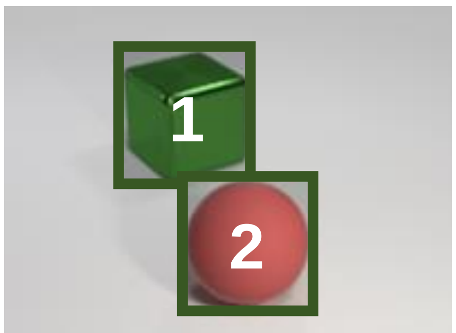
```
Filter(Red)
Query(Shape)
```

Program

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision



Scene Parsing

ID	Color	Shape	Material
1	Green	Cube	Metal
2	Red	Sphere	Rubber

Symbolic Reasoning

Language

Q: What's the **shape** of the **red** object?

Semantic Parsing

```
Filter(Red)
Query(Shape)
```

Program

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision

Scene Parsing

ID	Color	Shape	Material
1	Green	Cube	Metal
2	Red	Sphere	Rubber

Language

Q: What's the **shape** of the **red** object?

Semantic Parsing

```
Filter(Red)
Query(Shape)
```

Program

Symbolic Reasoning

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision

Scene Parsing

ID	Color	Shape	Material
1	Green	Cube	Metal
2	*Red*	Sphere	Rubber

Symbolic Reasoning

Language

Q: What's the **shape** of the **red** object?

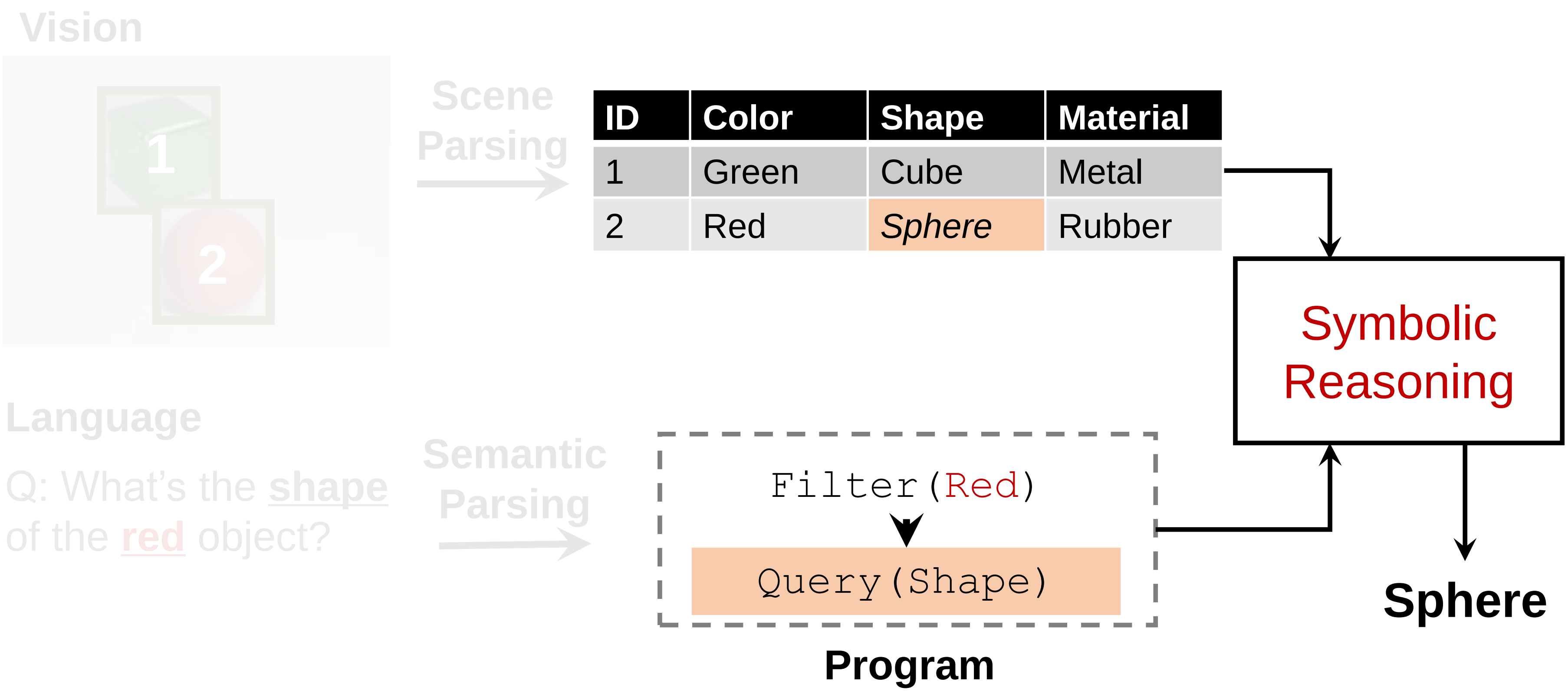
Semantic Parsing

```
Filter(Red)
Query(Shape)
```

Program

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

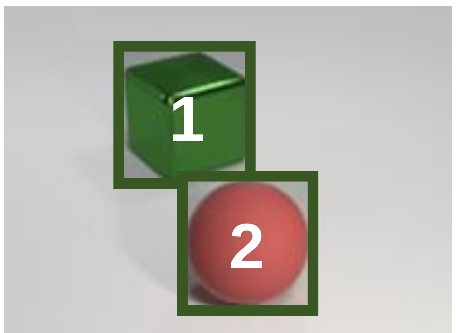


ID	Color	Shape	Material
1	Green	Cube	Metal
2	Red	*Sphere*	Rubber

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision



Scene Parsing

Concept Annotation

Language

Q: What's the **shape** of the **red** object?

Semantic Parsing

Program Annotation

Symbolic Reasoning

Sphere

NS-VQA [Yi et al. 2018]

Incorporate Concepts in Visual Reasoning

Vision

Scene Parsing

Concept Annotation?

Language

Q: Are the **animals** **grazing**?

VQA [Agrawal et al., 2015]

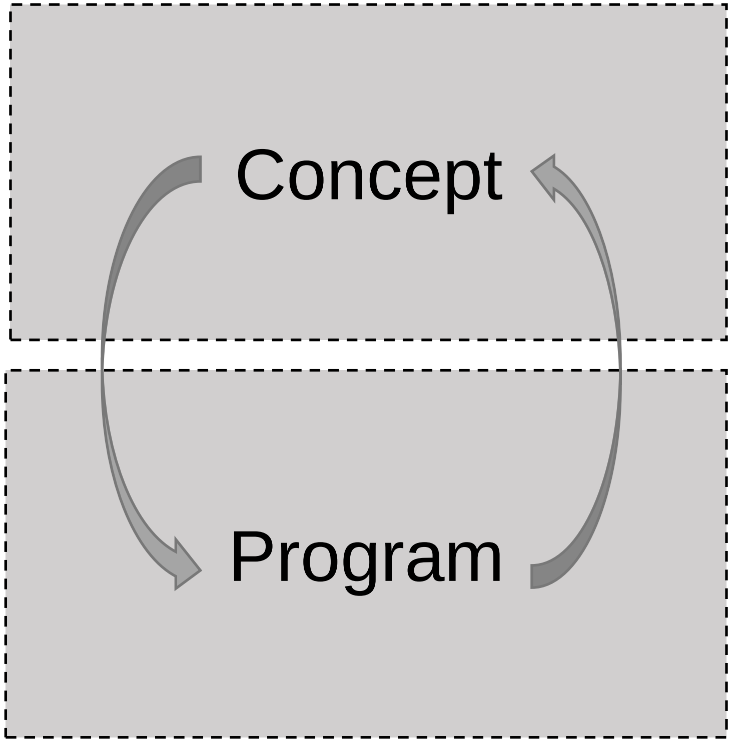
Semantic Parsing

Program Annotation?

Idea: Joint Learning of Concepts and Semantic Parsing

Vision

Scene Parsing



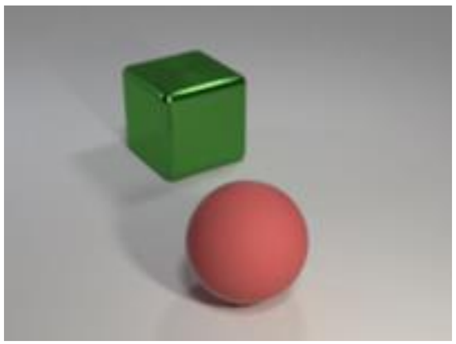
Language

Q: Are the **animals grazing**?

Semantic Parsing

VQA [Agrawal et al., 2015]

Idea: Joint Learning of Concepts and Semantic Parsing



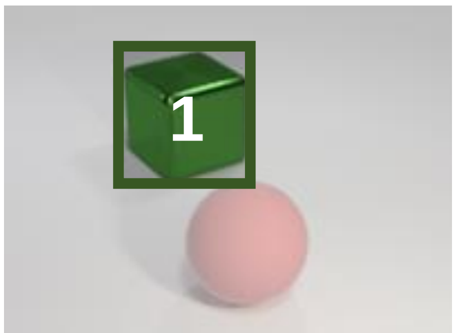
Object Detection

Feature Extraction

Visual Representation

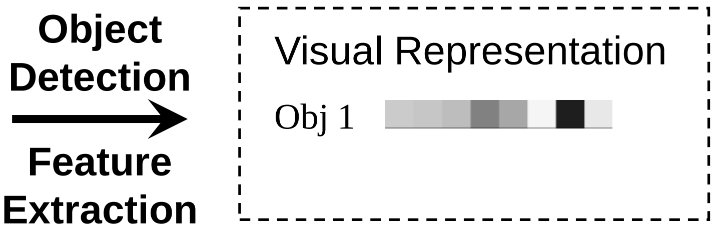
Q: What's the **shape** of the **red** object?

Idea: Joint Learning of Concepts and Semantic Parsing



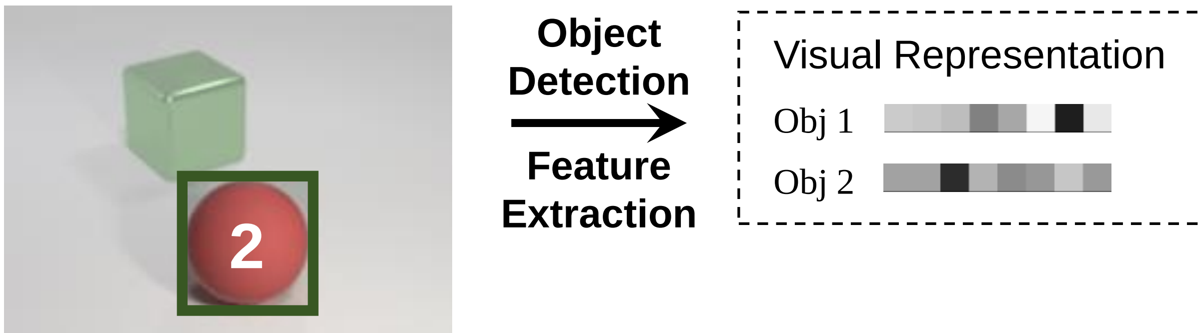
Object Detection

Feature Extraction



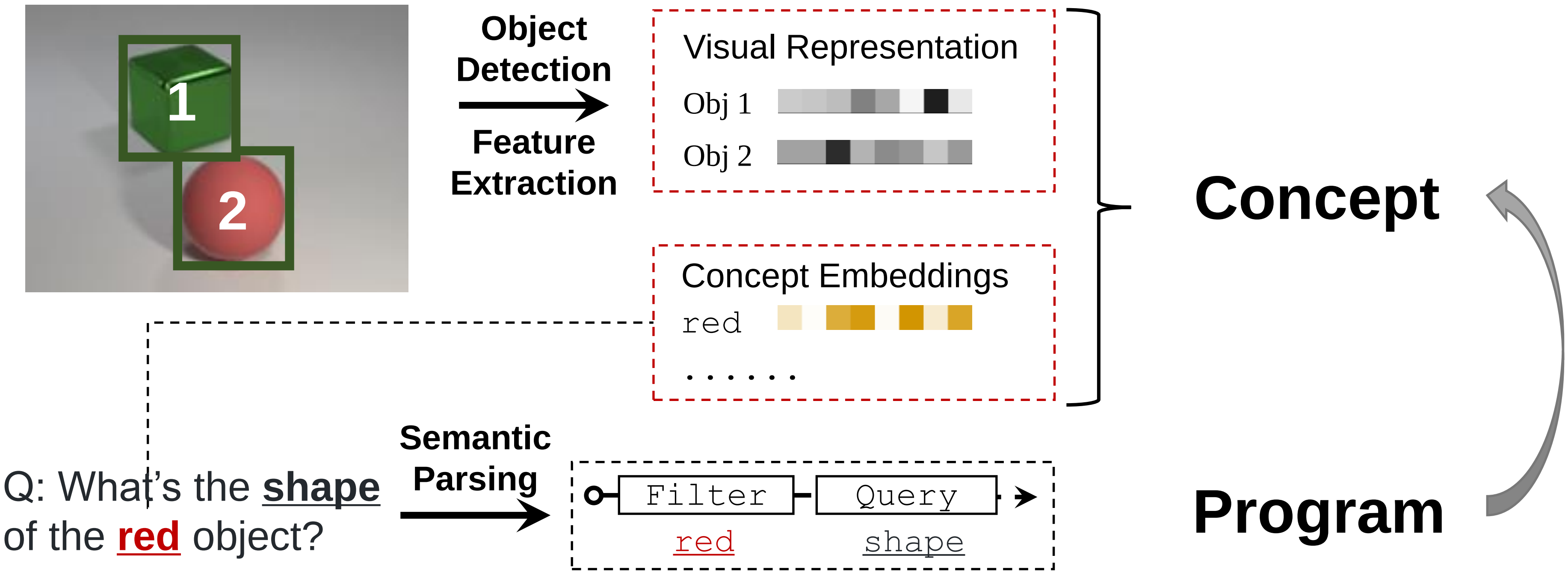
Q: What's the **shape** of the **red** object?

Idea: Joint Learning of Concepts and Semantic Parsing

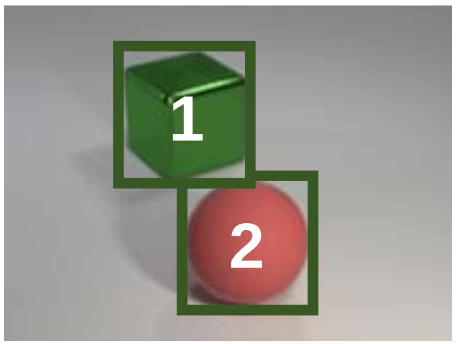


Q: What’s the **shape**
of the **red** object?

Idea: Joint Learning of Concepts and Semantic Parsing



Neuro-Symbolic Reasoning



Q: What's the **shape** of the **red** object?

Visual Representation

Obj 1

Obj 2

Concept Embeddings

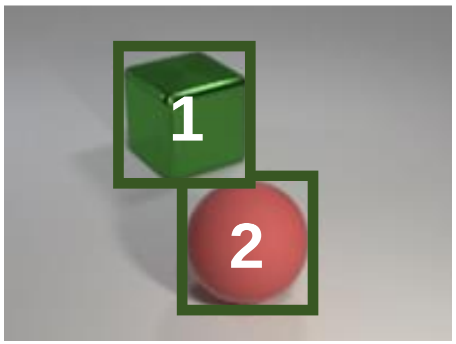
red

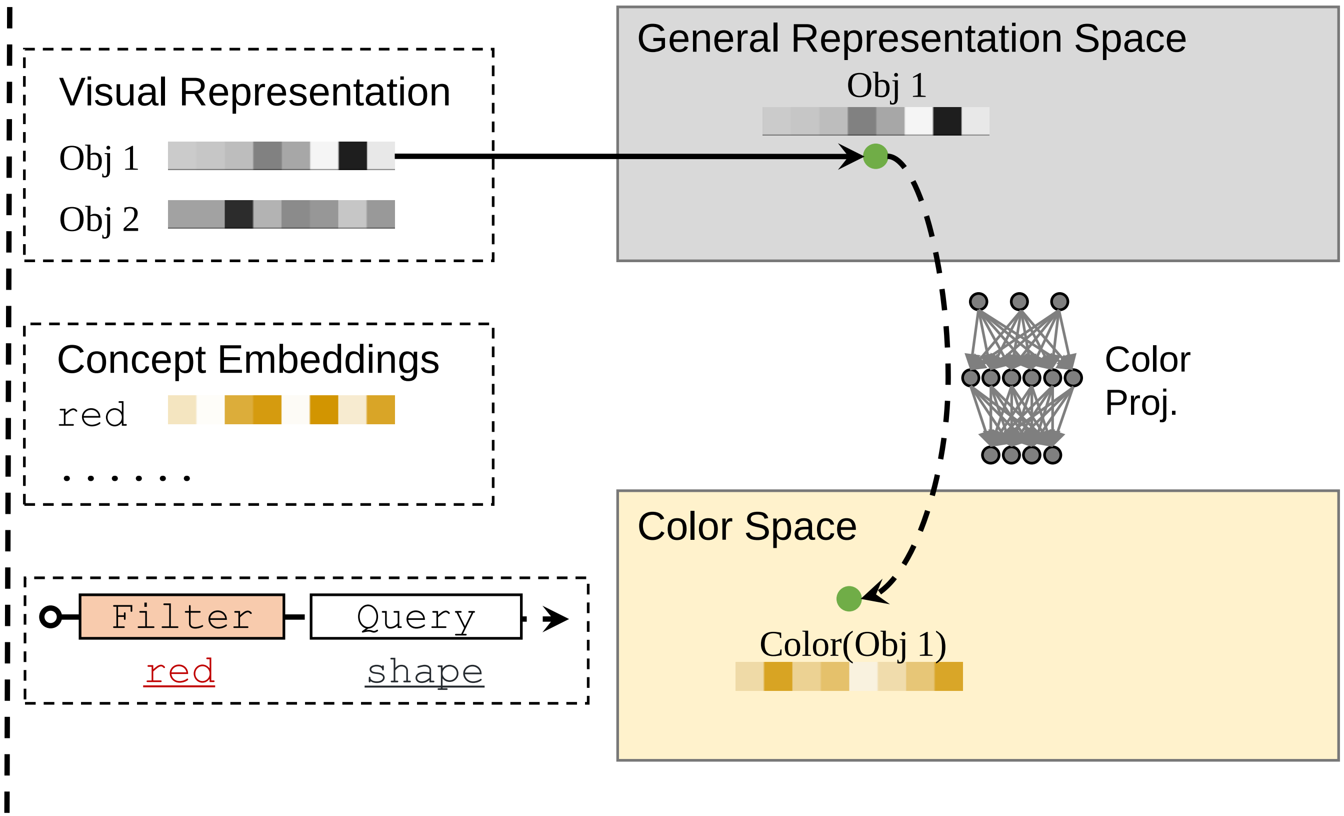
......

Filter → Query

red shape

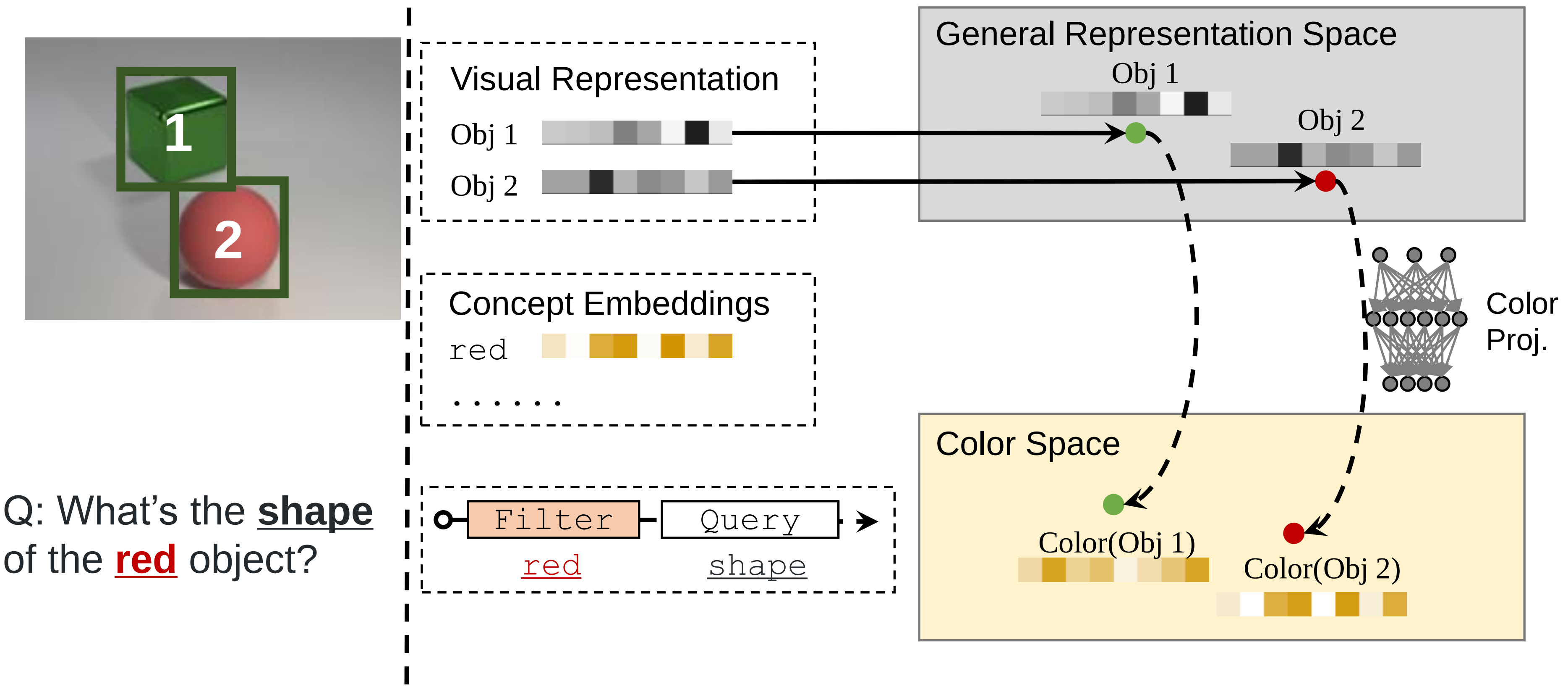
Neuro-Symbolic Reasoning



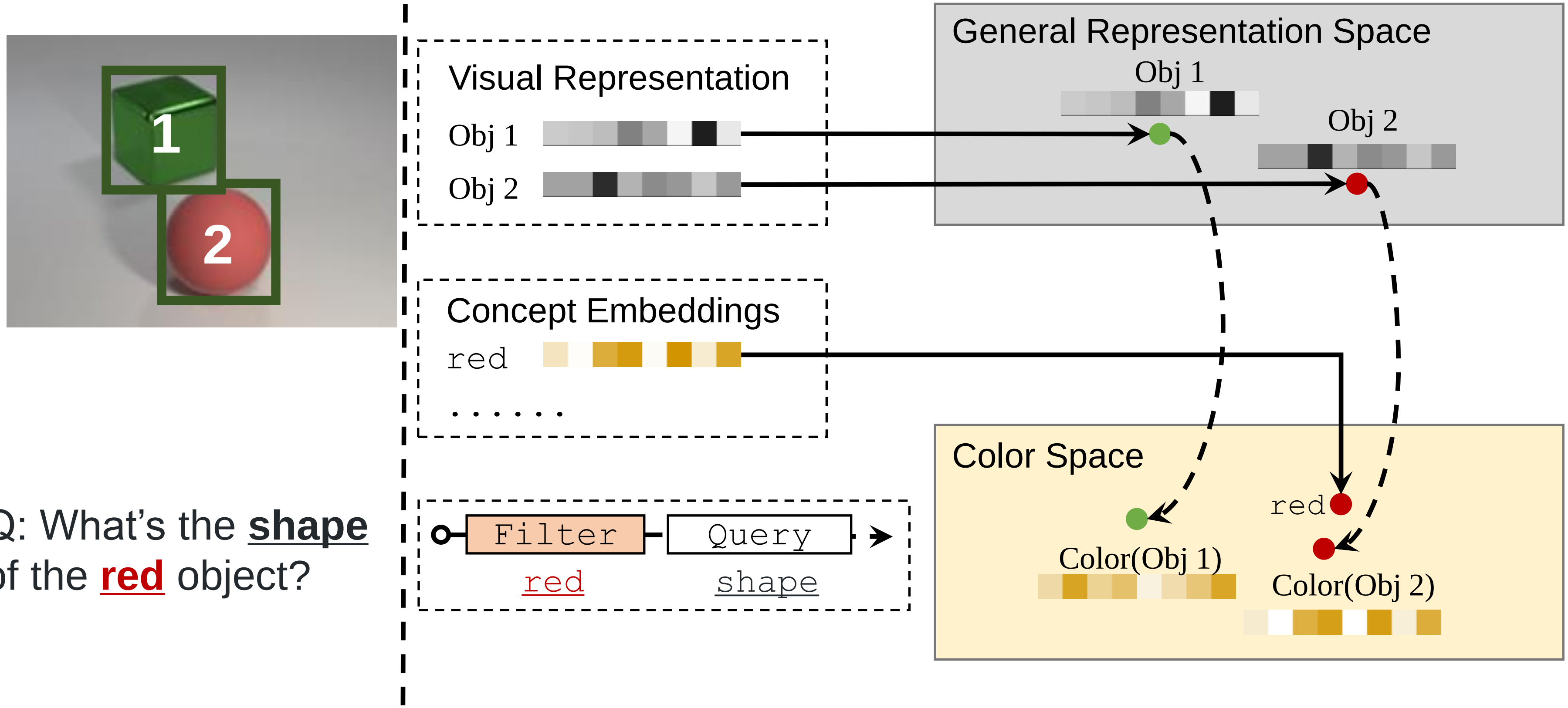


Q: What's the **shape** of the **red** object?

Neuro-Symbolic Reasoning

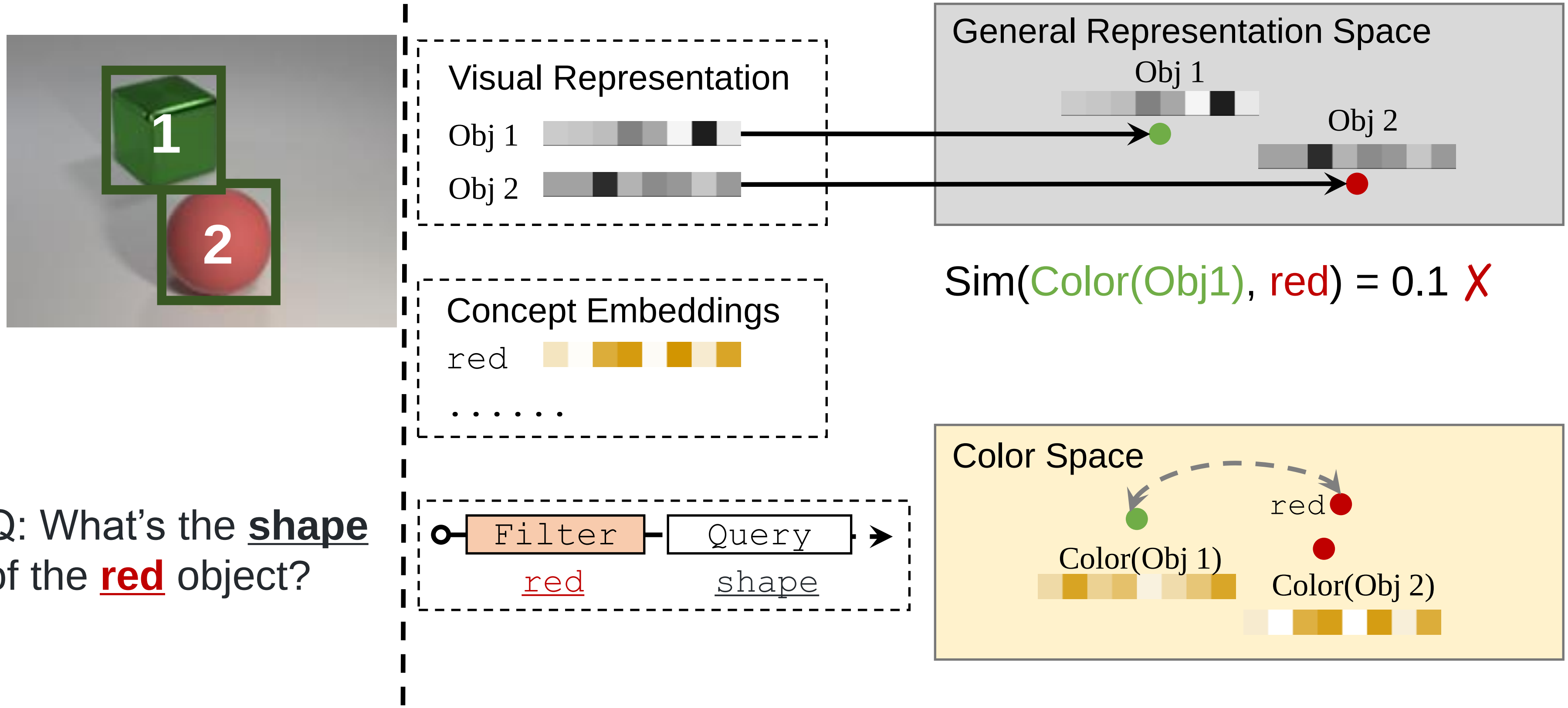


Neuro-Symbolic Reasoning



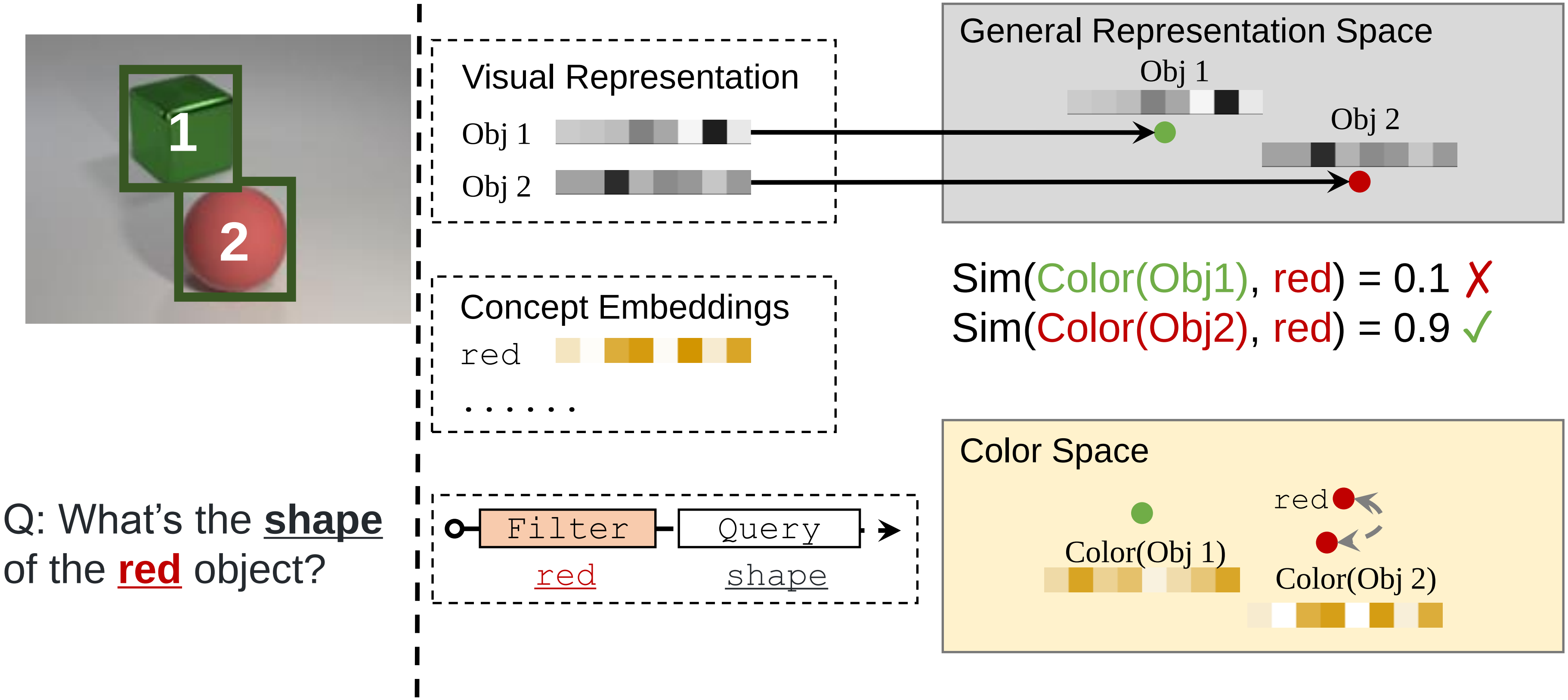
Q: What's the **shape** of the **red** object?

Neuro-Symbolic Reasoning

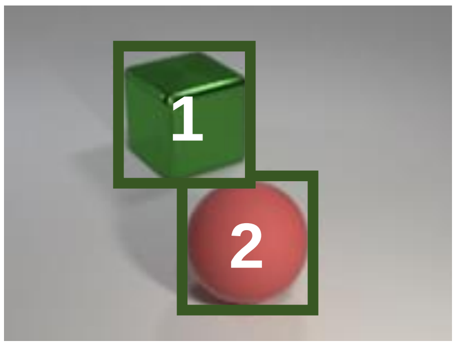


Q: What's the **shape** of the **red** object?

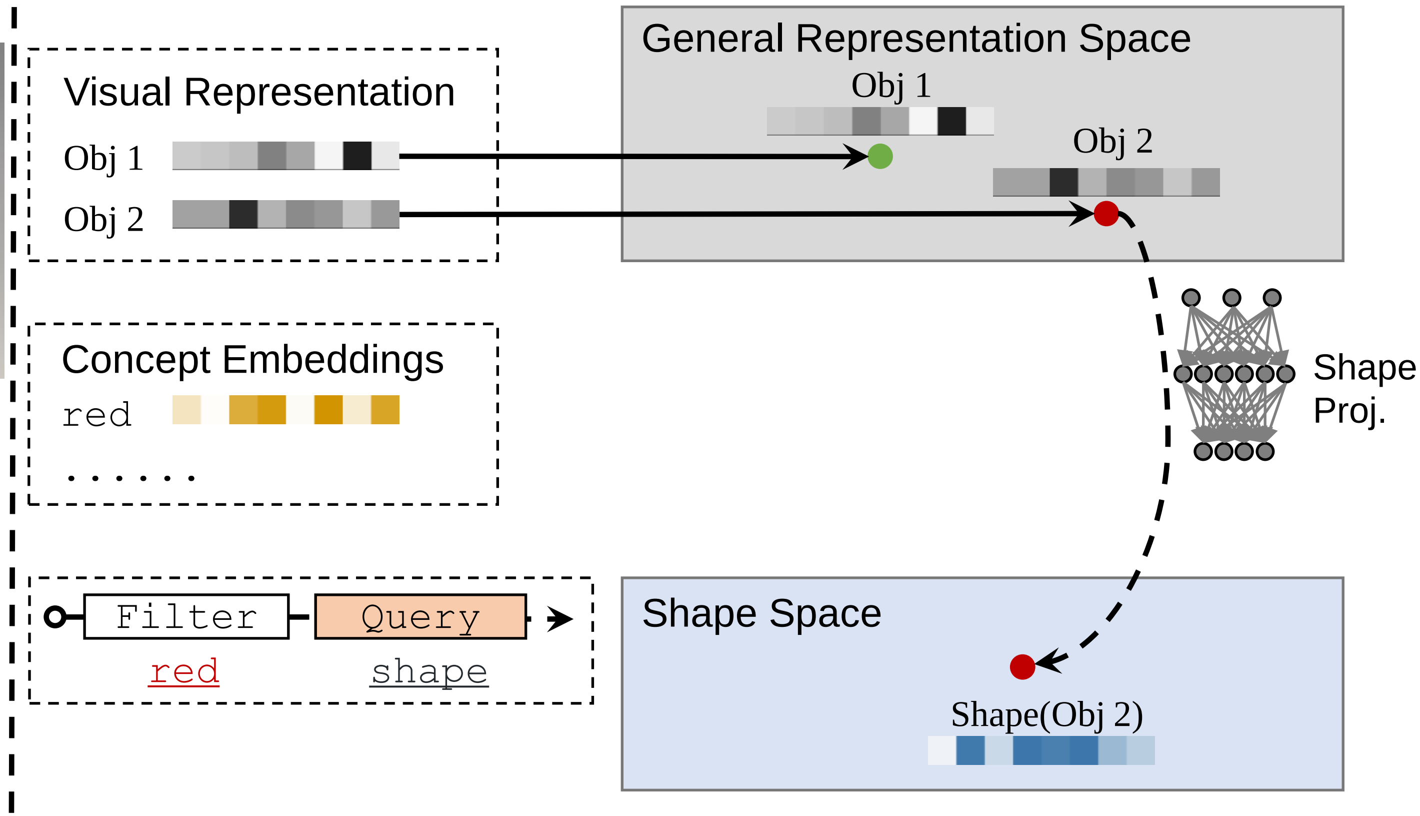
Neuro-Symbolic Reasoning



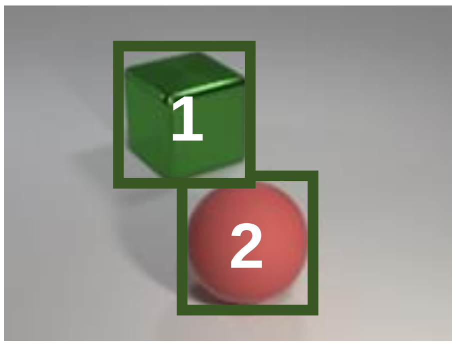
Neuro-Symbolic Reasoning



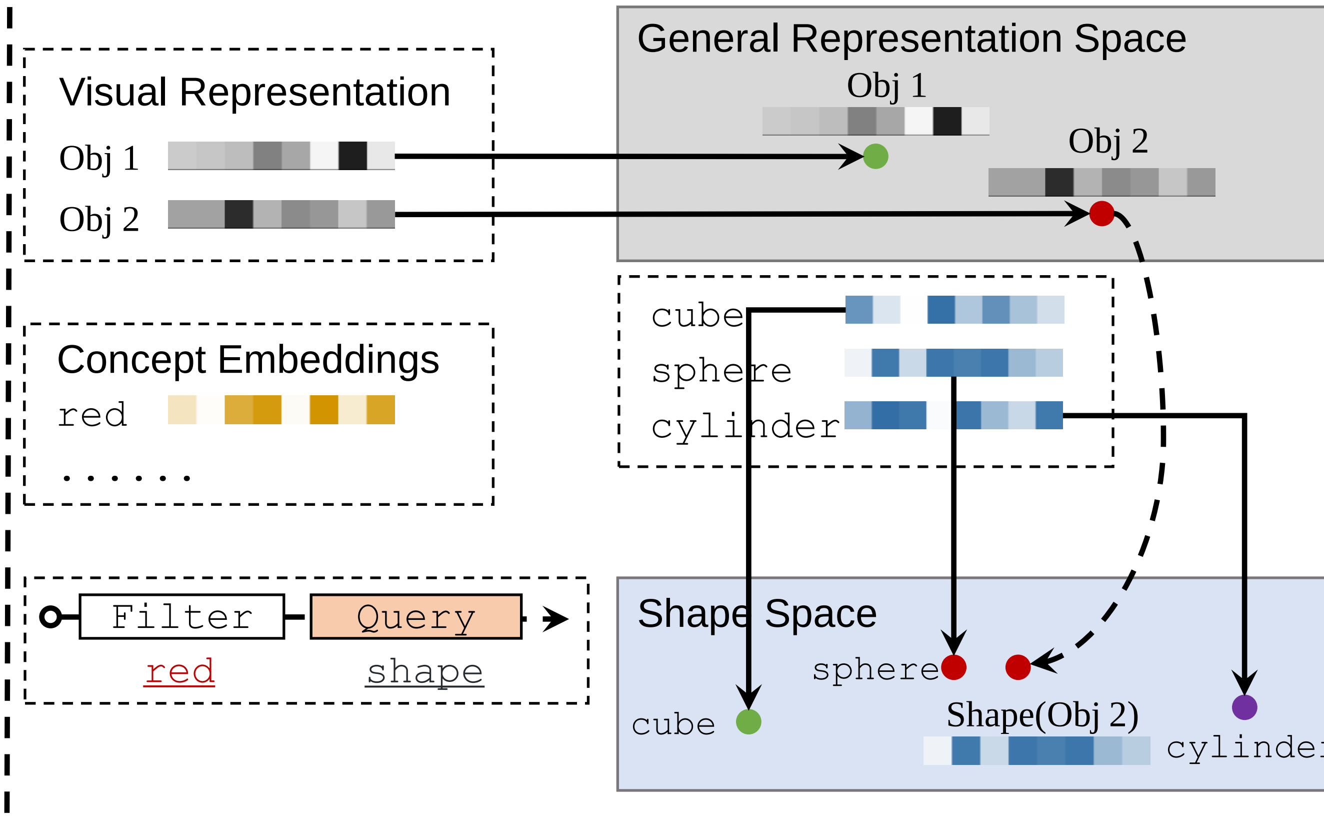
Q: What's the **shape** of the **red** object?



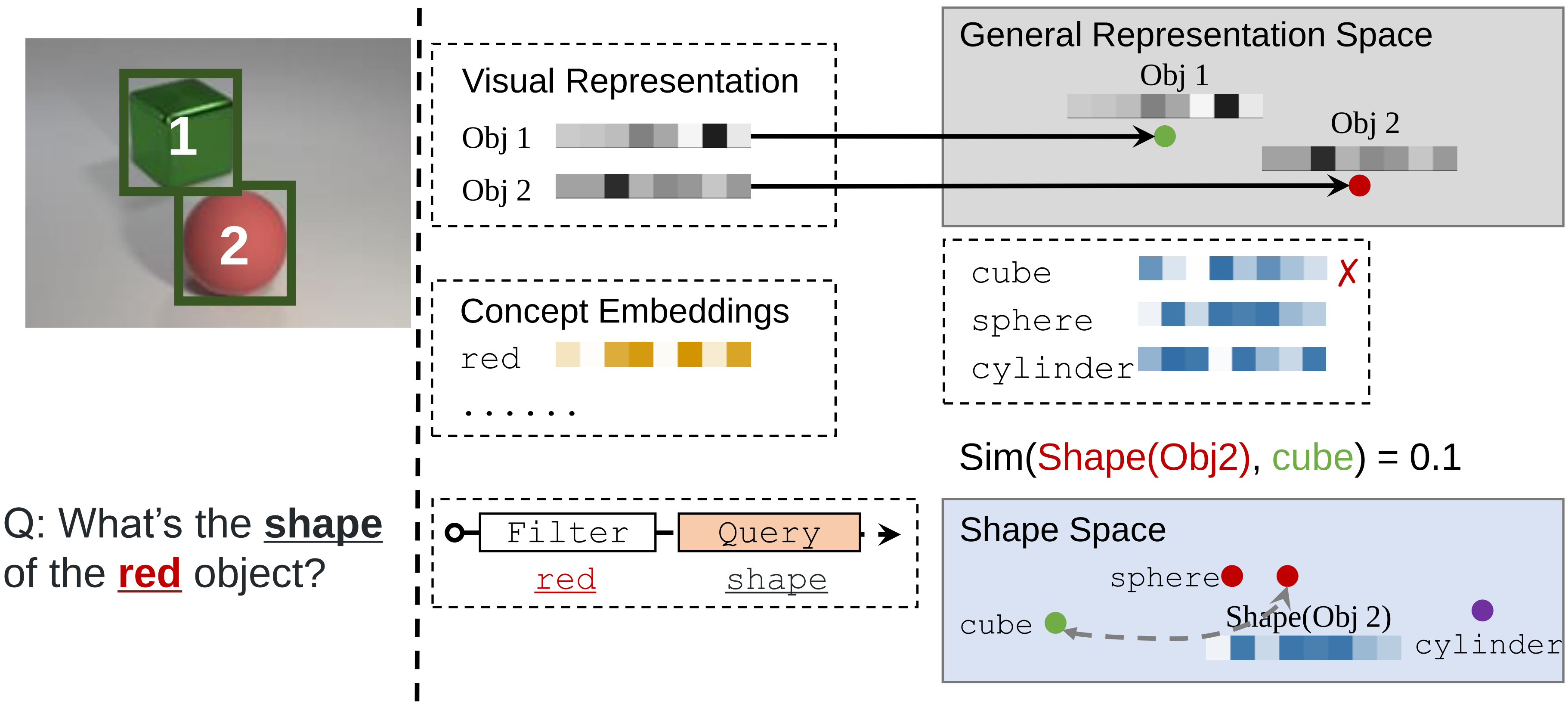
Neuro-Symbolic Reasoning



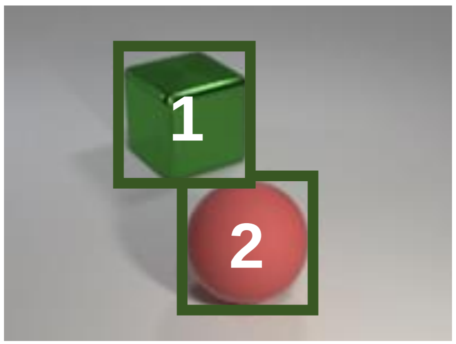
Q: What's the **shape** of the **red** object?

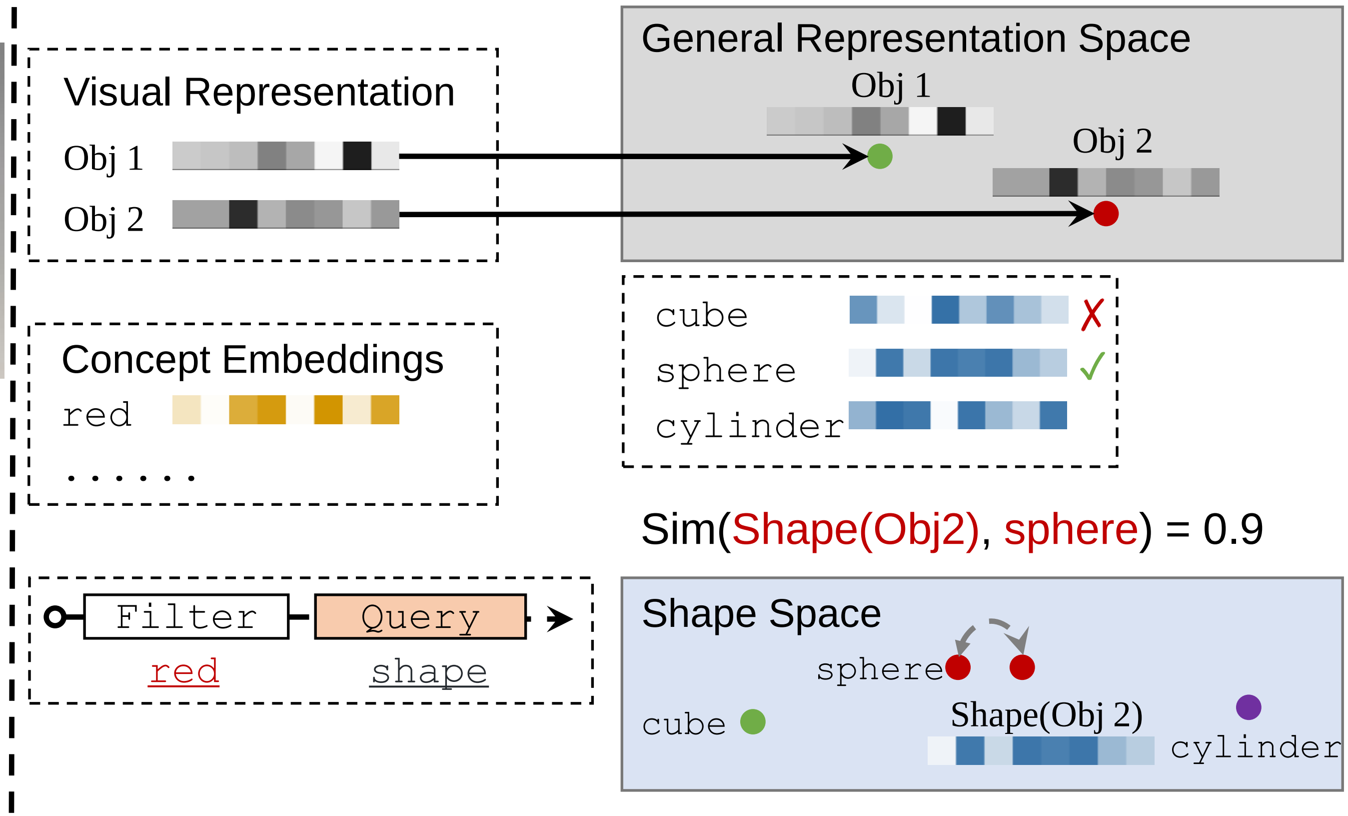


Neuro-Symbolic Reasoning



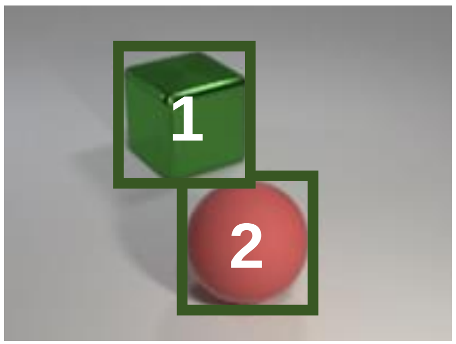
Neuro-Symbolic Reasoning



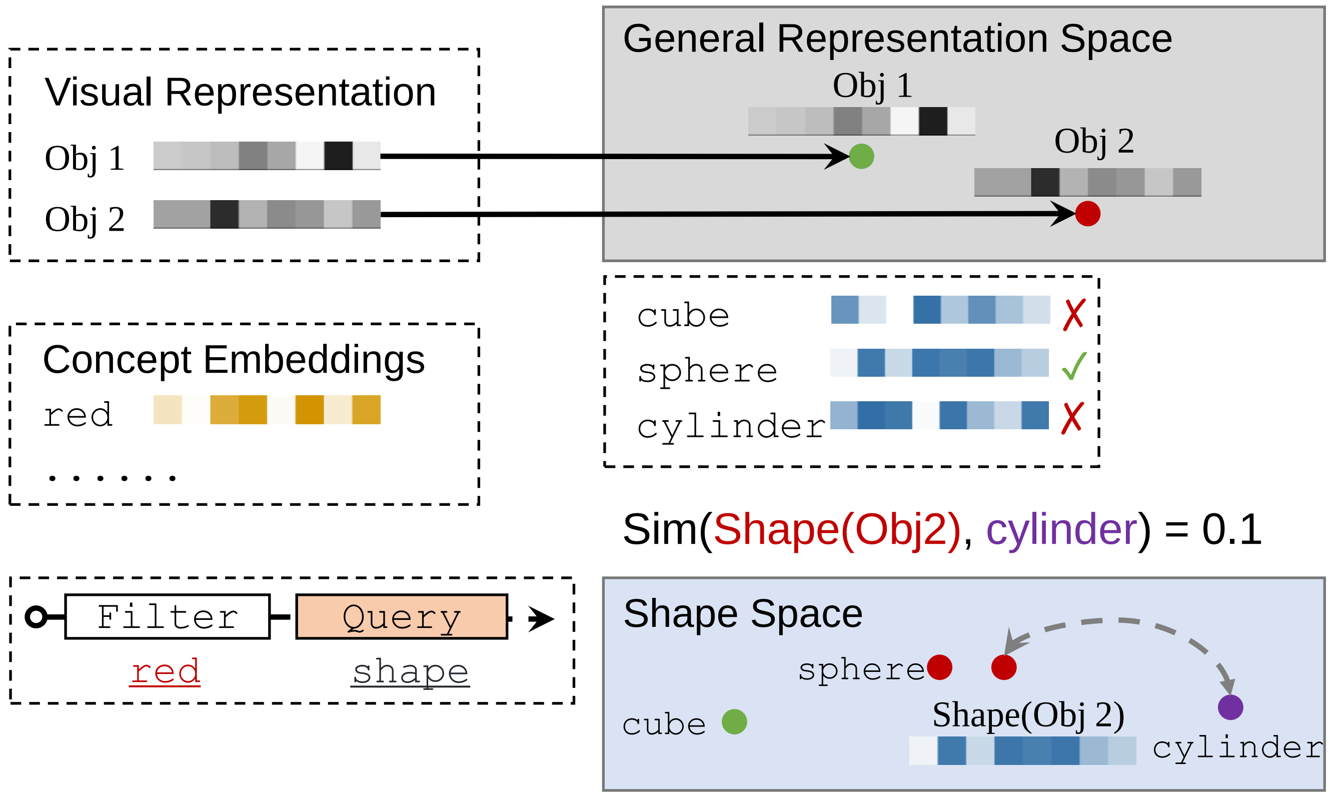


Q: What's the **shape** of the **red** object?

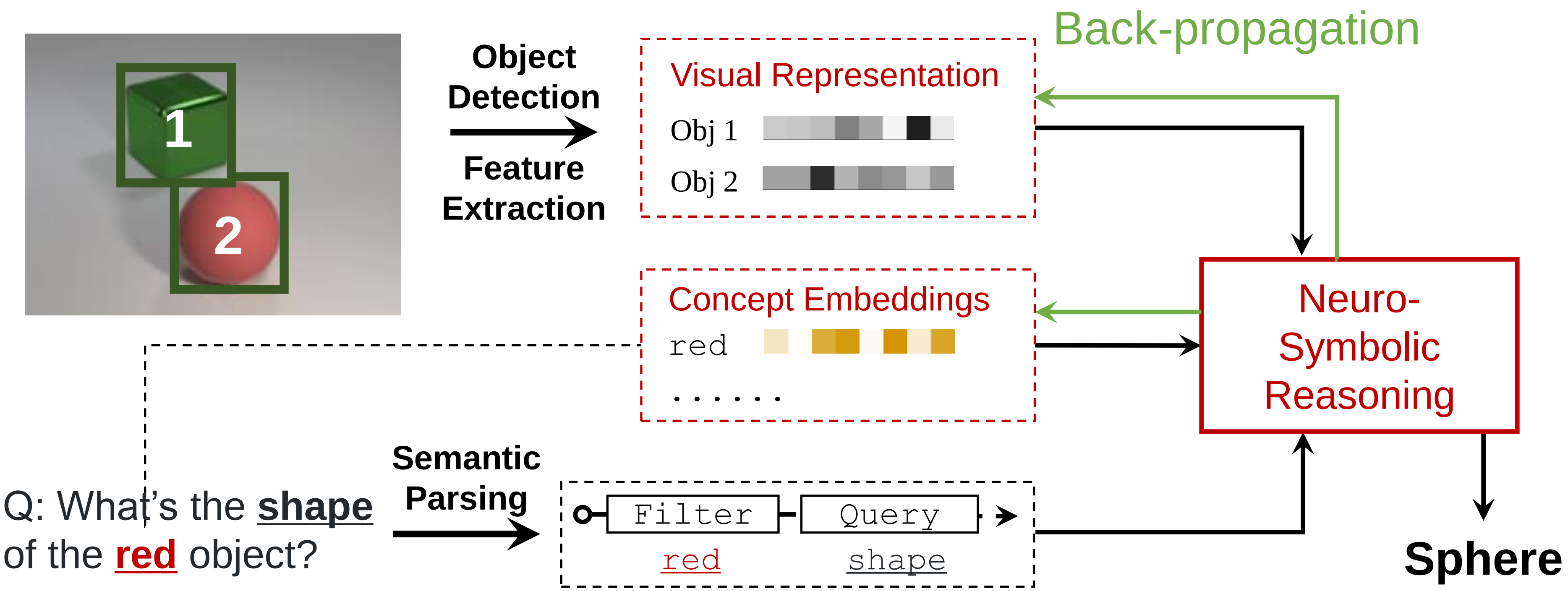
Neuro-Symbolic Reasoning



Q: What's the **shape** of the **red** object?

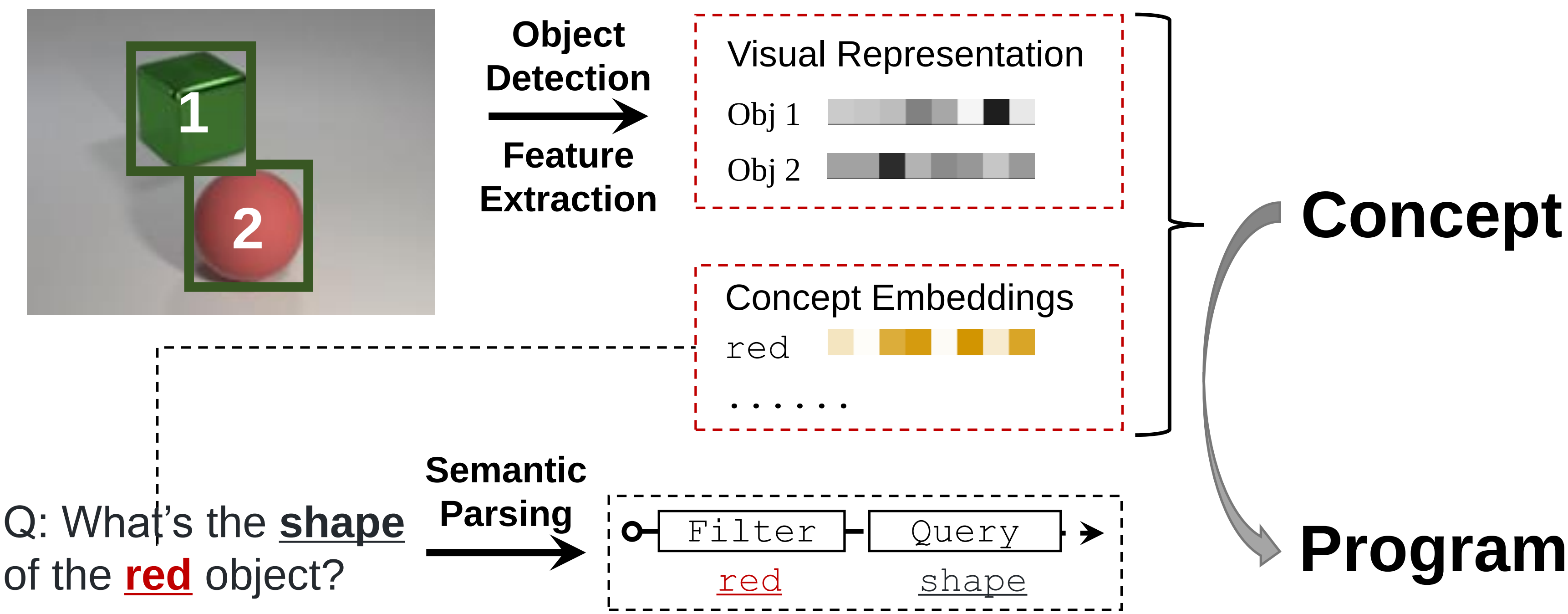


Neuro-Symbolic Reasoning



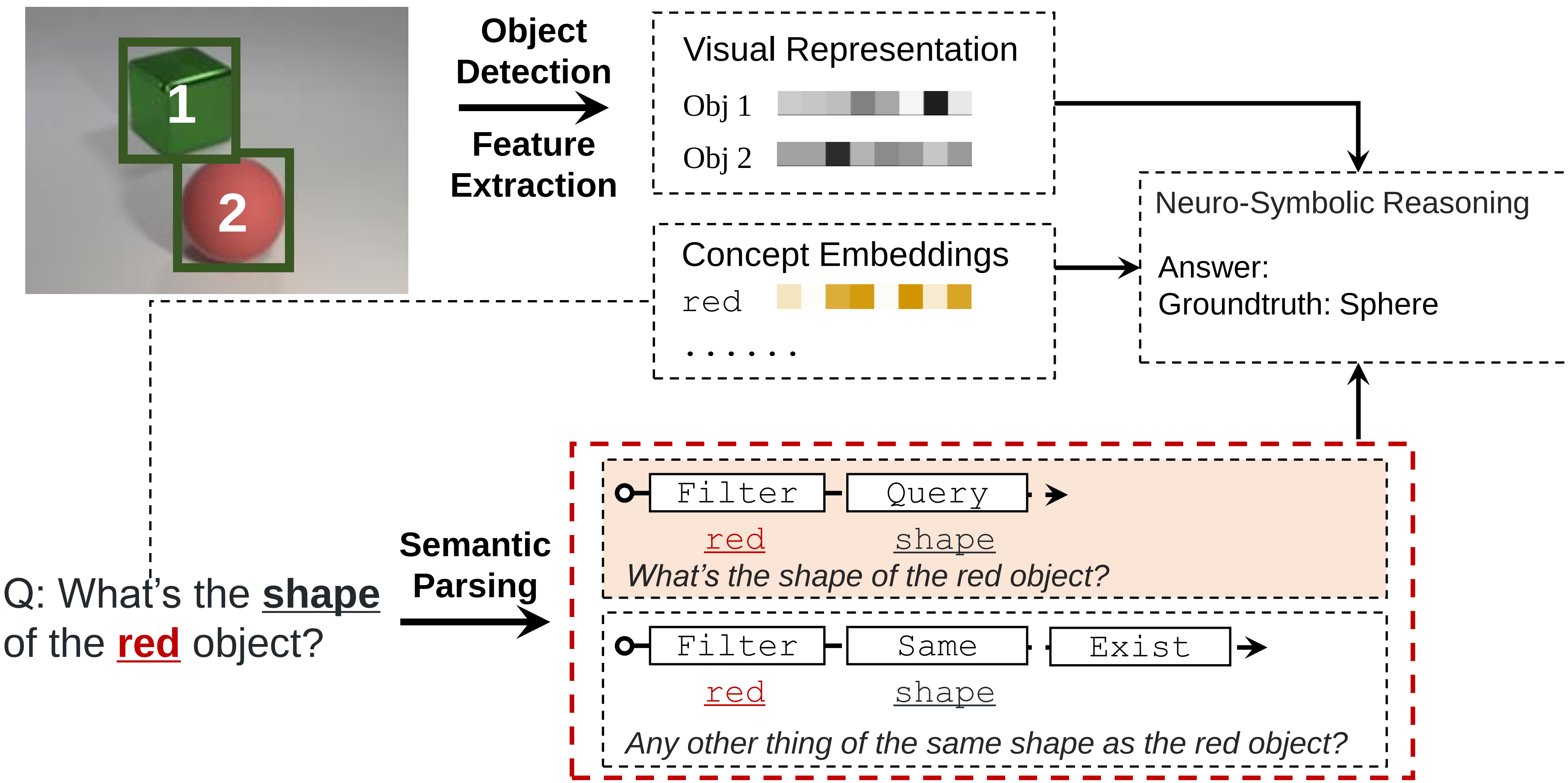
Jiayuan Mao

Idea: Joint Learning of Concepts and Semantic Parsing

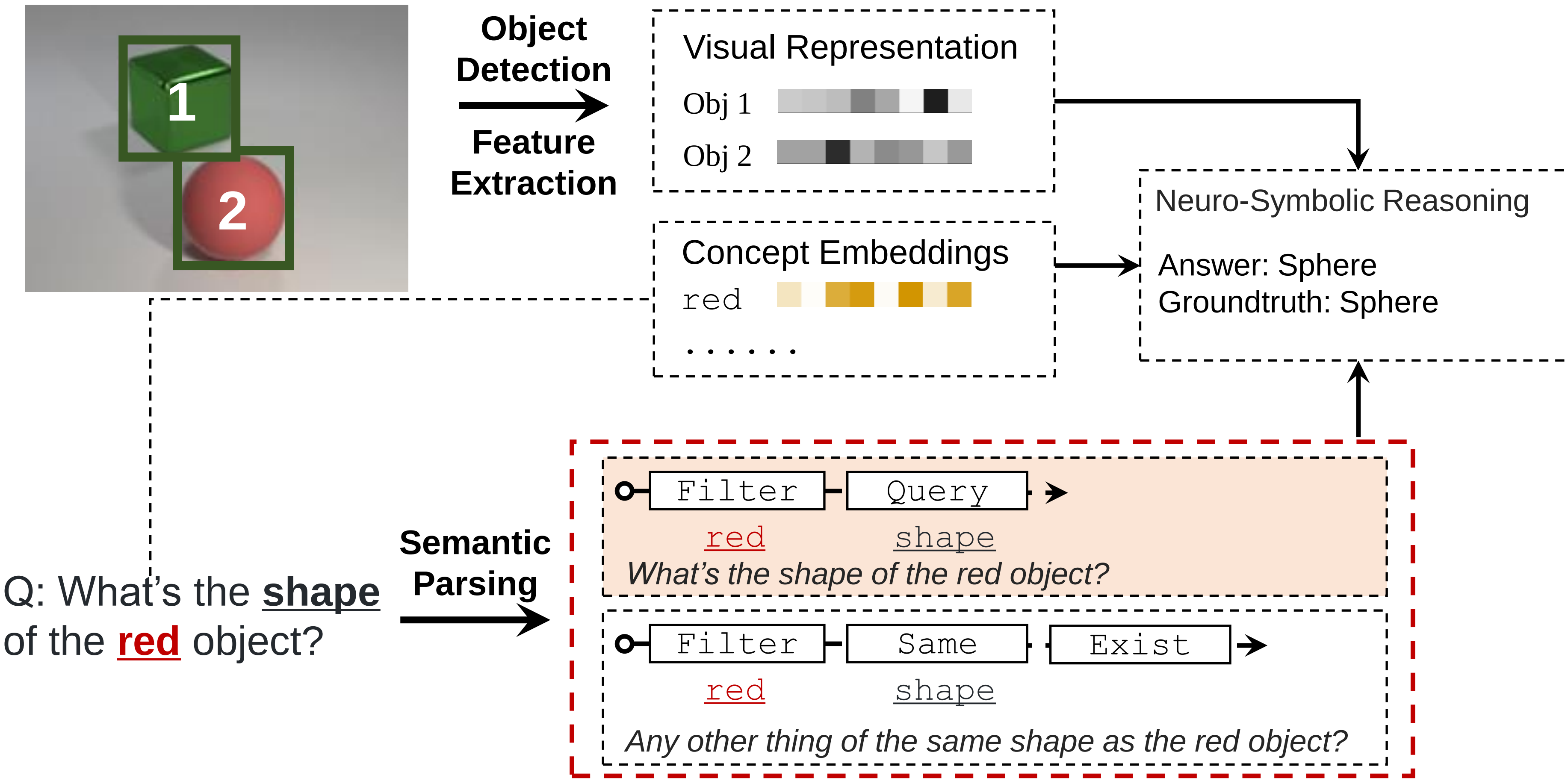


Jiayuan Mao

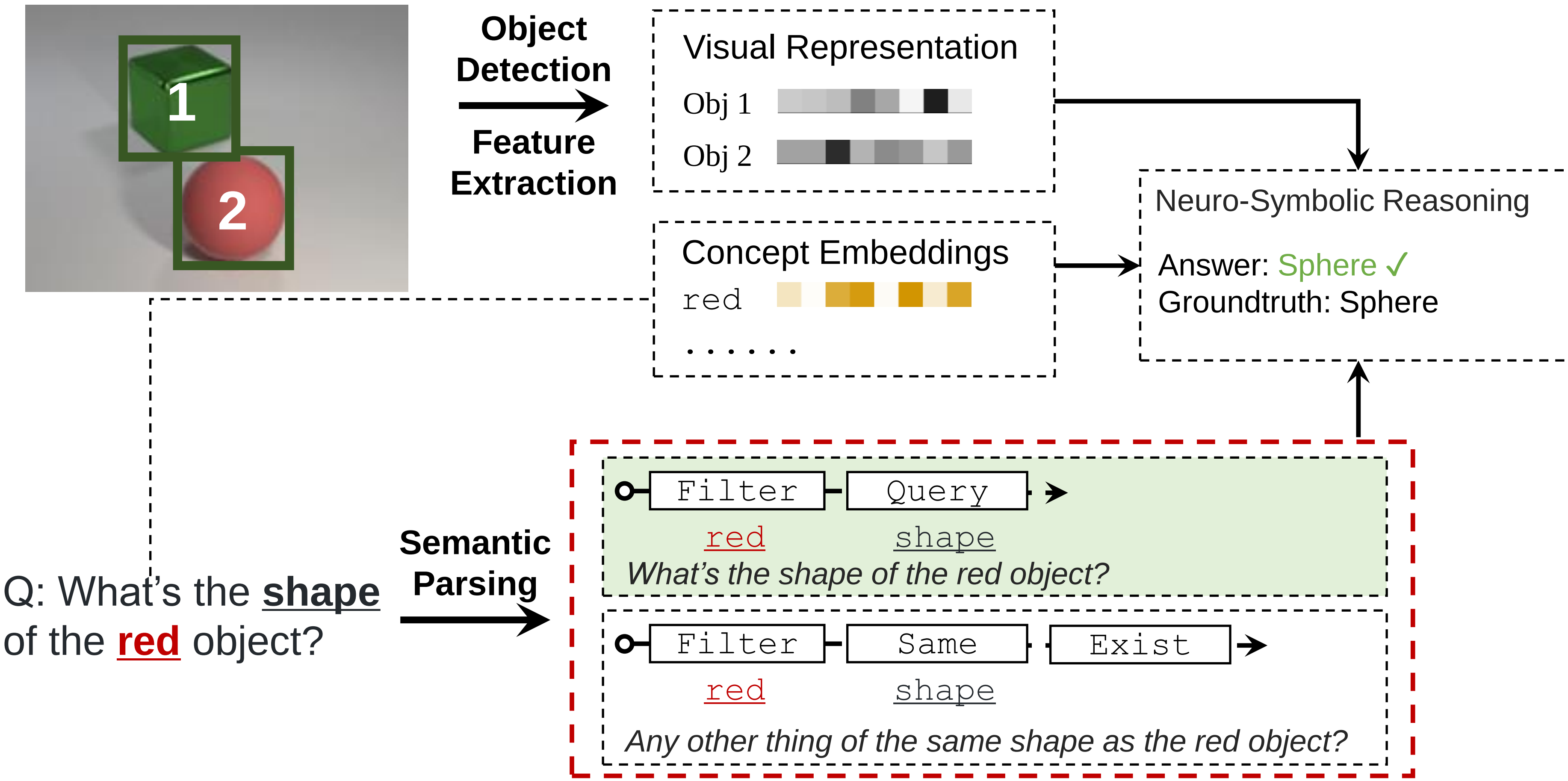
Concepts Facilitate Parsing New Sentences



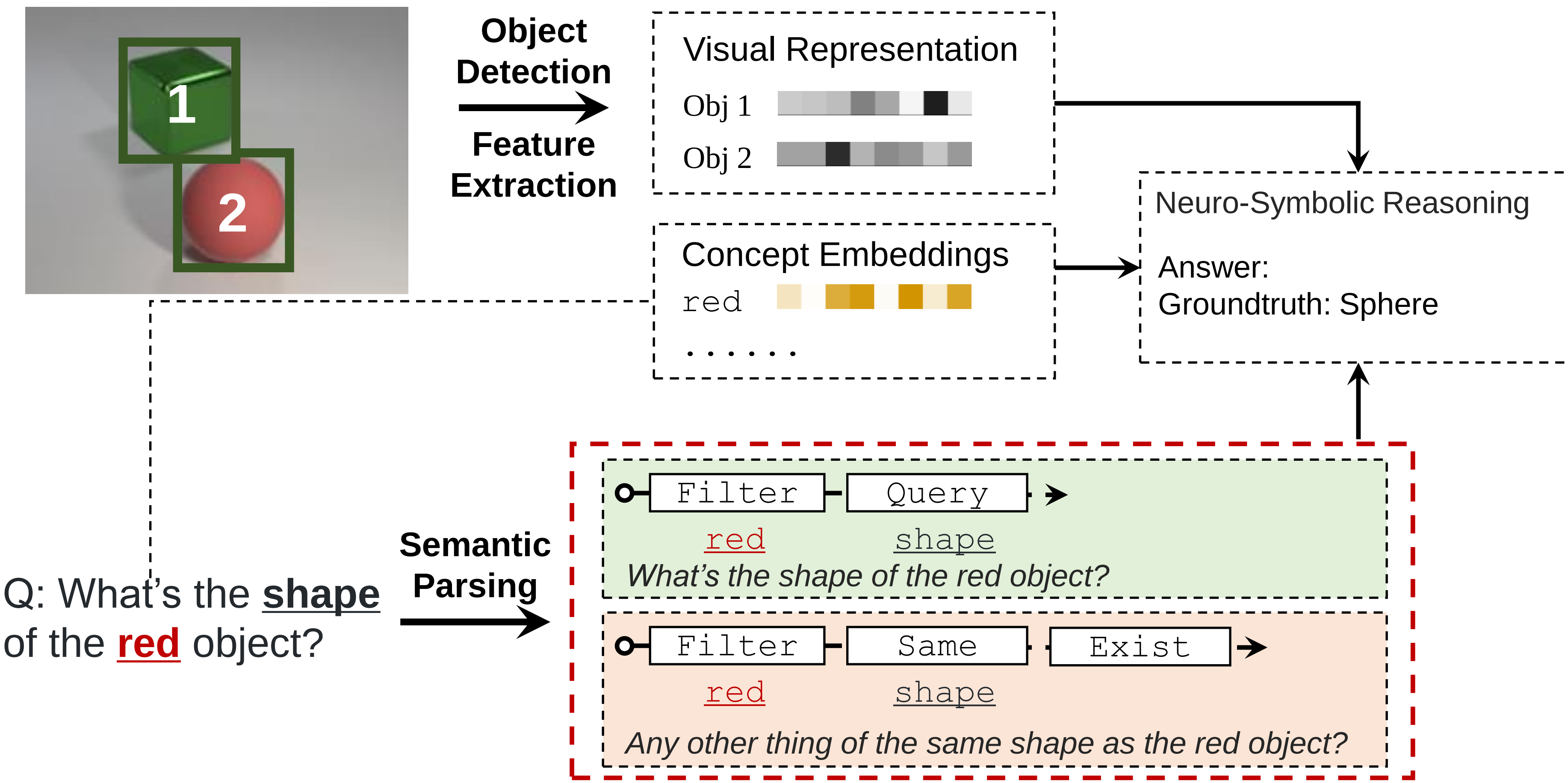
Concepts Facilitate Parsing New Sentences



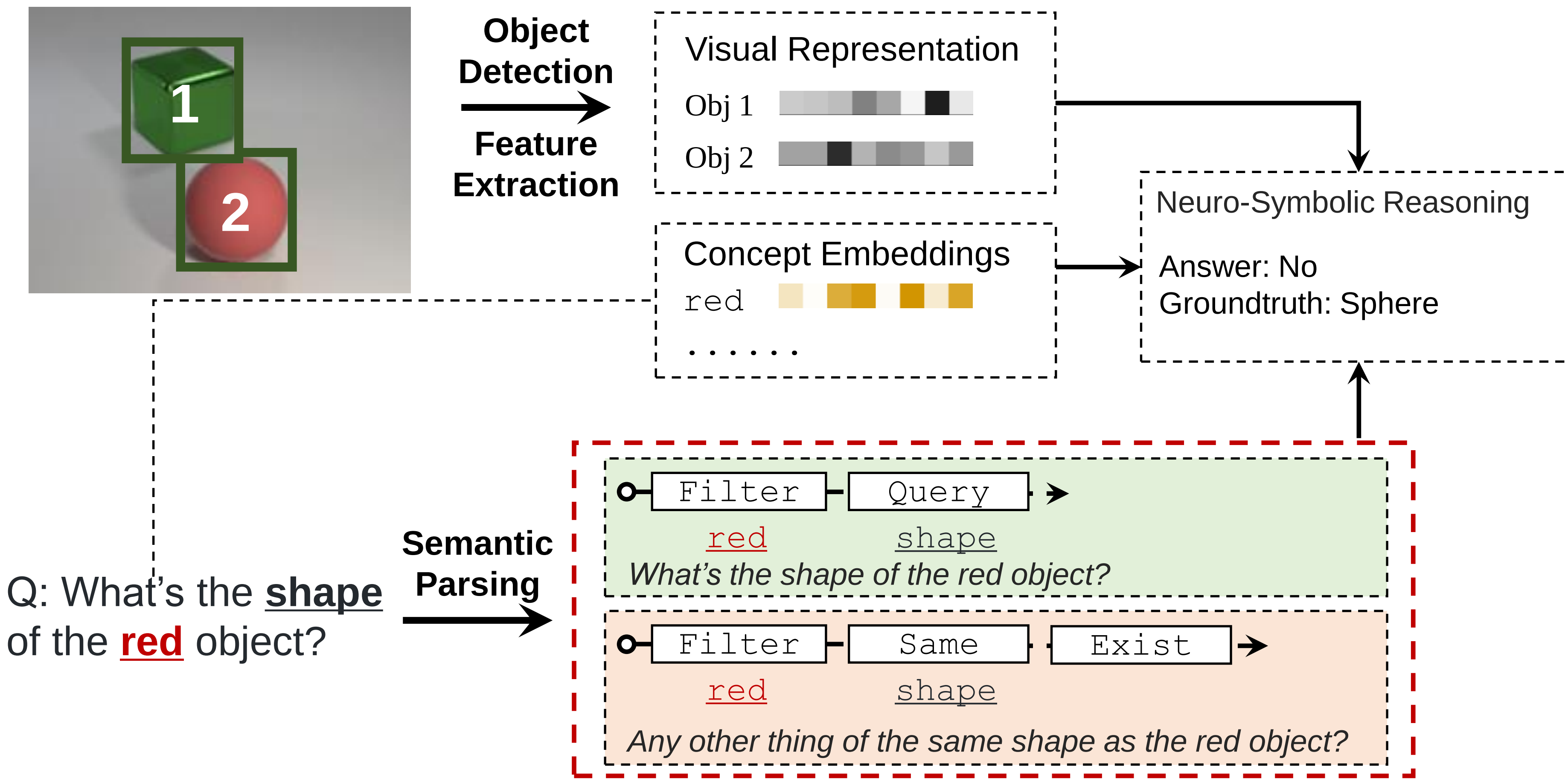
Concepts Facilitate Parsing New Sentences



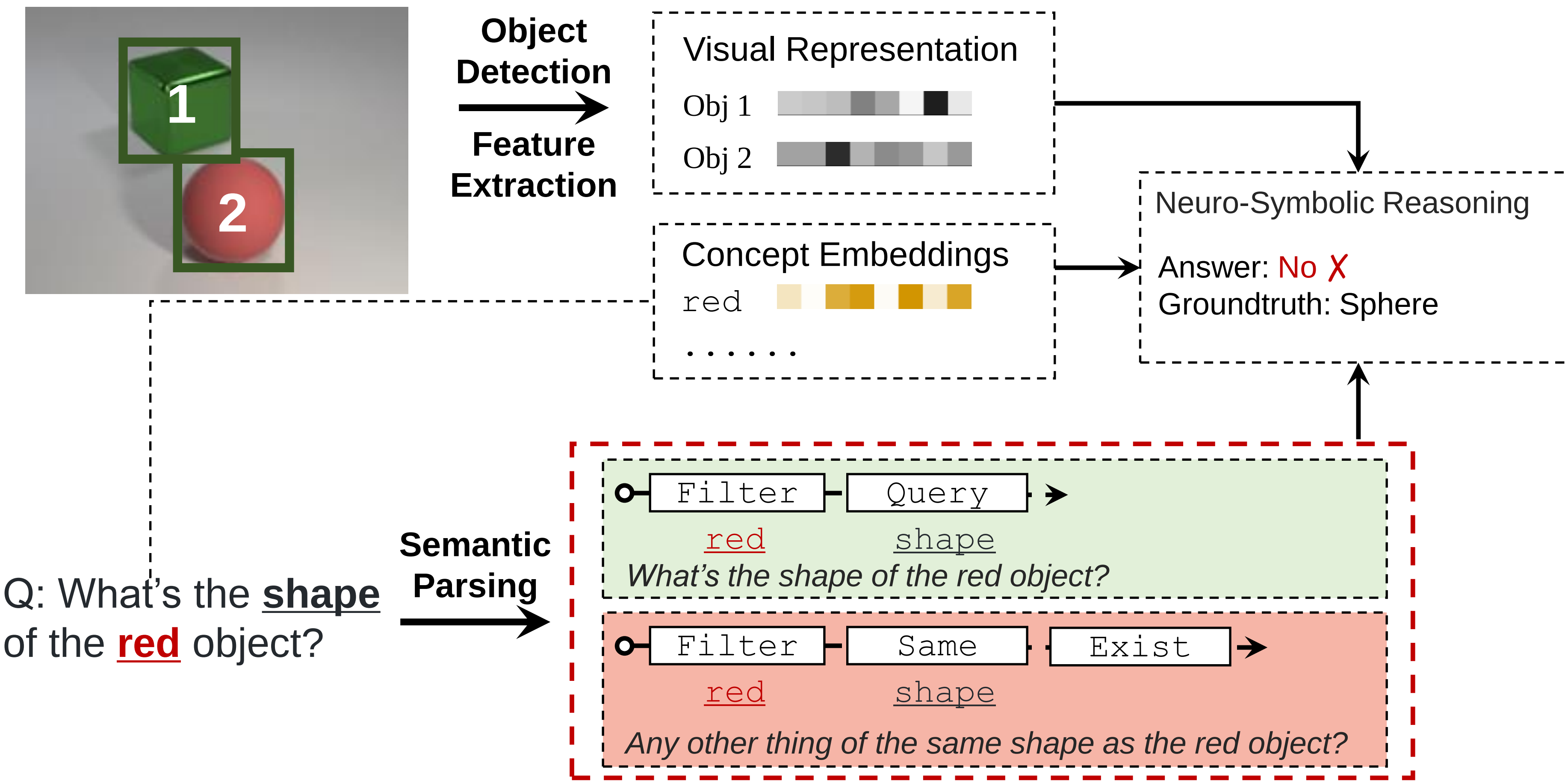
Concepts Facilitate Parsing New Sentences



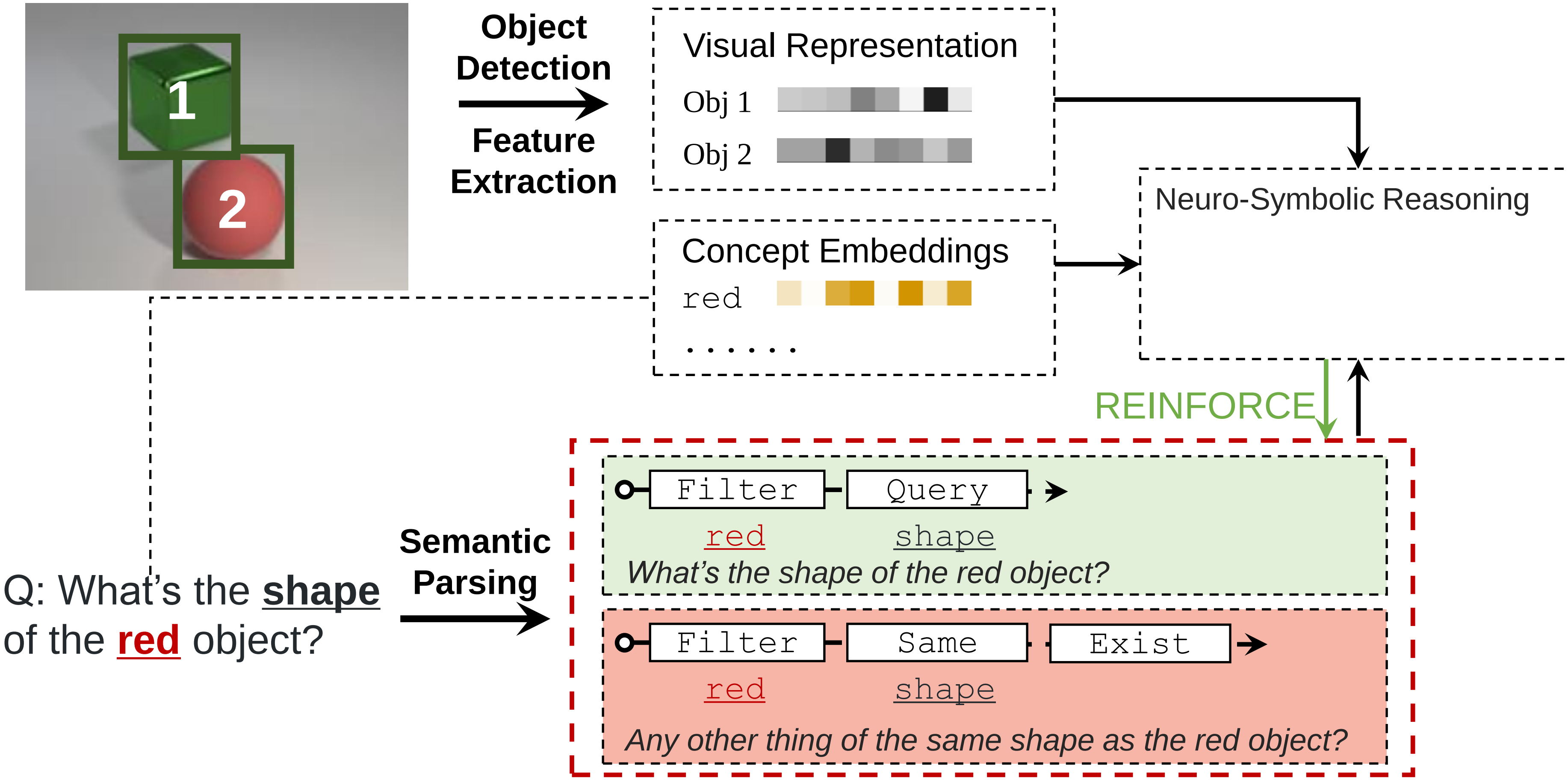
Concepts Facilitate Parsing New Sentences



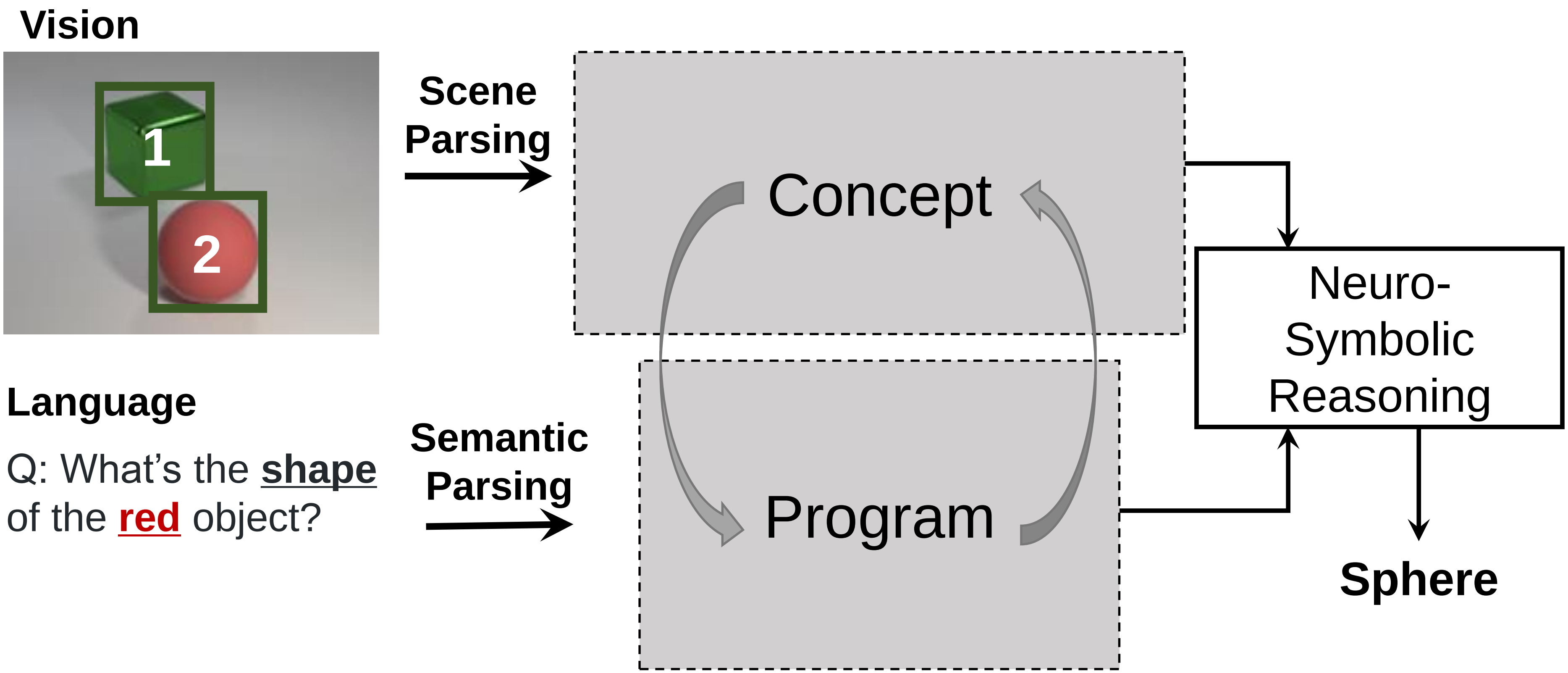
Concepts Facilitate Parsing New Sentences



Concepts Facilitate Parsing New Sentences

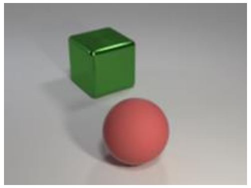


Idea: Joint Learning of Concepts and Semantic Parsing



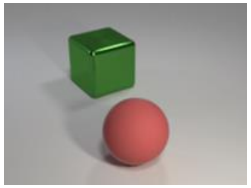
Curriculum Learning

Lesson1: Object-based questions.



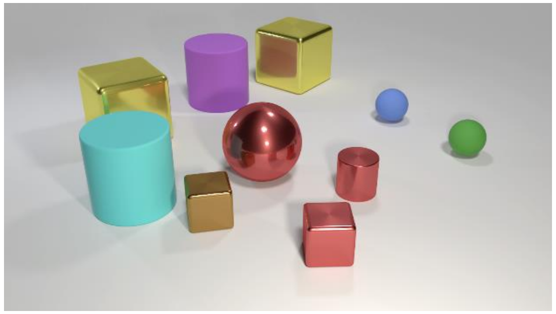
Q: What is the shape of the red object?
A: Sphere.

Lesson2: Relational questions.



Q: Is the green cube behind the red sphere?
A: Yes

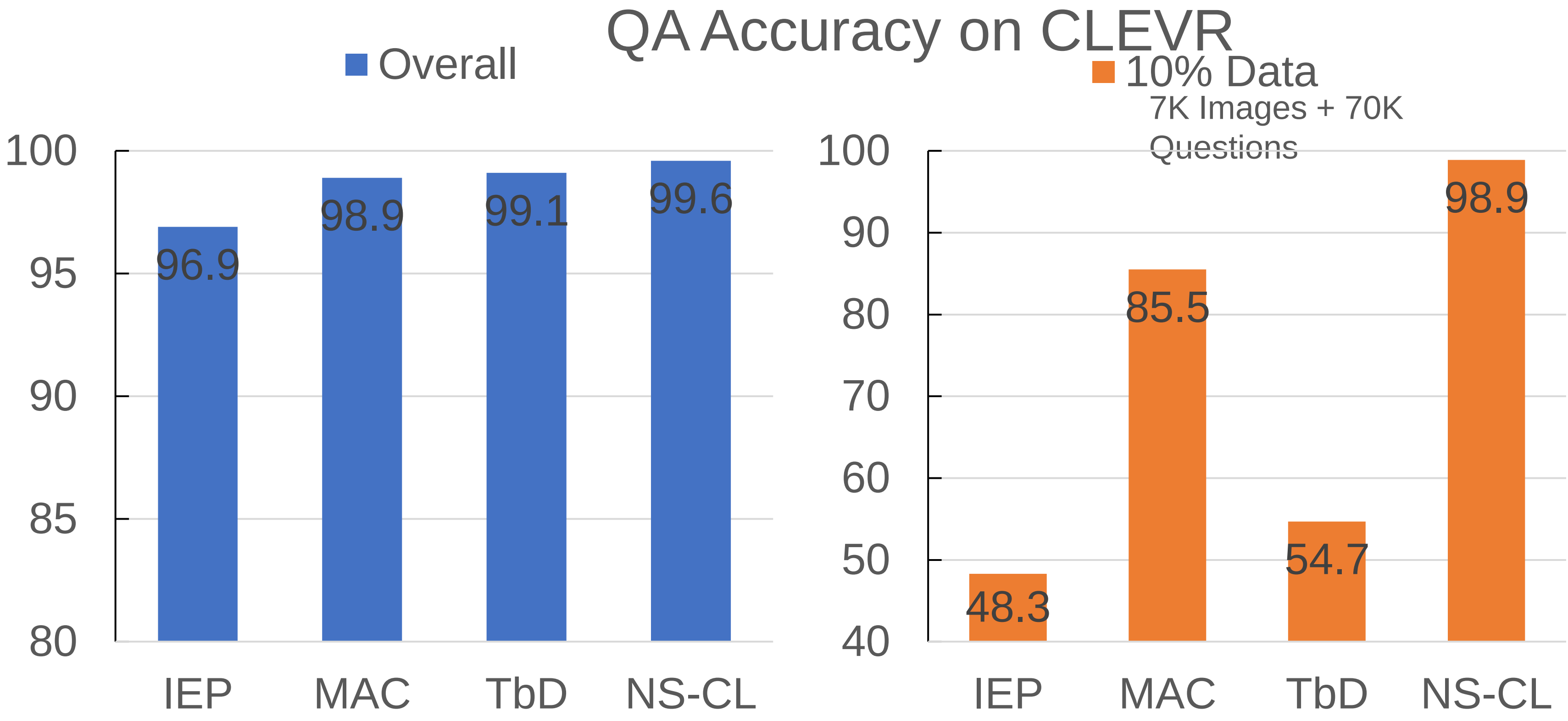
Lesson3: complex scenes, complex questions



Q: Does the big matte object behind the big sphere have the same color as the cylinder left of the small brown cube?
A: No.

High Accuracy and Data Efficiency

IEP [Johnson et al. 2017]
MAC [Hudson & Manning, 2018]
TbD [Mascharka et al. 2018]
NS-CL [Ours]



Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

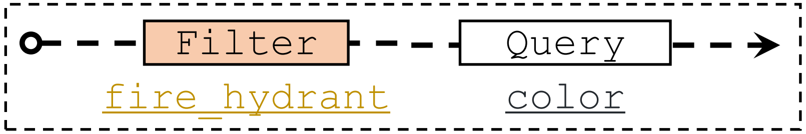
Q: What **color** is the **fire hydrant**?

Filter → Query
fire_hydrant color

Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

Q: What **color** is the **fire hydrant**?



Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

Q: What **color** is the **fire hydrant**?

Filter → Query

fire_hydrant   color

A: Yellow

Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

Q: What **color** is the **fire hydrant**?

`Filter` → `Query`
fire_hydrant   color

A: Yellow

Q: How many **zebras** are there?

`Filter` → `Count`
zebra

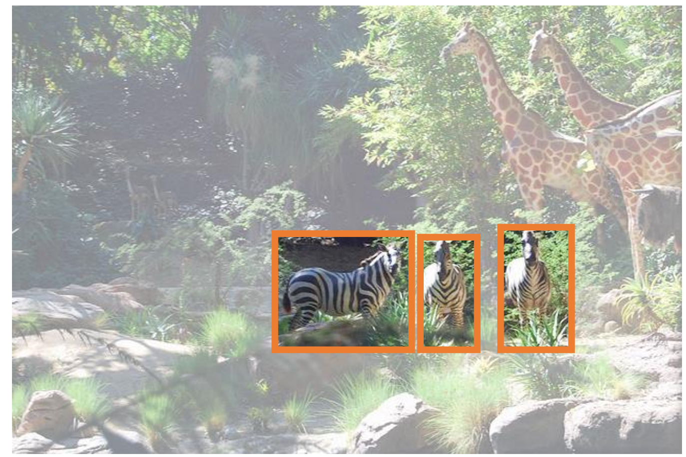
Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

Q: What **color** is the **fire hydrant**?

Filter → Query
fire_hydrant color

A: Yellow



Q: How many **zebras** are there?

Filter → Count
zebra

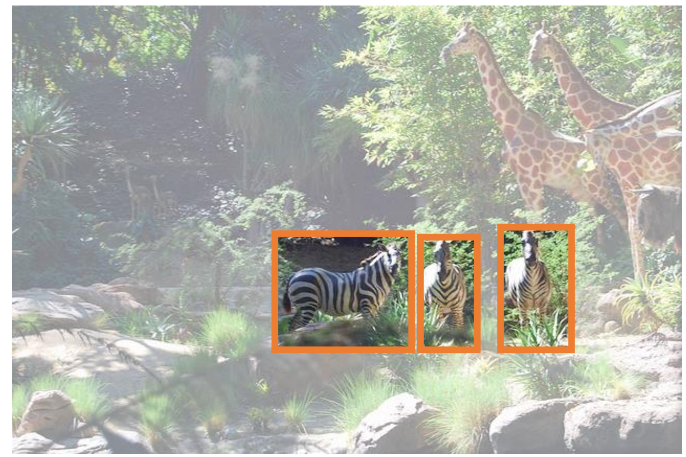
Application in Real-World Scenarios

VQA [Agrawal et al., 2015]
VQS [Gan et al., 2017]

Q: What **color** is the **fire hydrant**?

`Filter` → `Query`
fire_hydrant   color

A: Yellow



Q: How many **zebras** are there?

`Filter` → `Count`
zebra

A: 3

Graph-based Reasoning for VQA

Question: Which object in the image can be used to eat with?
Relation: UsedFor
Associated Fact: (Fork, UsedFor, Eat)
Answer Source: Image
Answer: Fork

Question: What do the animals in the image eat?
Relation: RelatedTo
Associated Fact: (Sheep, RelatedTo, Grass Eater)
Answer Source: Knowledge Base
Answer: Grass

Question: Which equipment in this image is used to hit baseball?
Relation: CapableOf
Associated Fact: (Baseball bat, CapableOf, Hit a baseball)
Answer Source: Image
Answer: Baseball bat

Fig. 1. The FVQA dataset expects methods to answer questions about images utilizing information from the image, as well as fact-based knowledge bases. Our method makes use of the image, and question text features, as well as high-level visual concepts extracted from the image in combination with a learned fact-ranking neural network. Our method is able to answer both visually grounded as well as fact based questions.

Narasimhan and Schwing, "Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering", ECCV 2018

Graph-based Reasoning for VQA

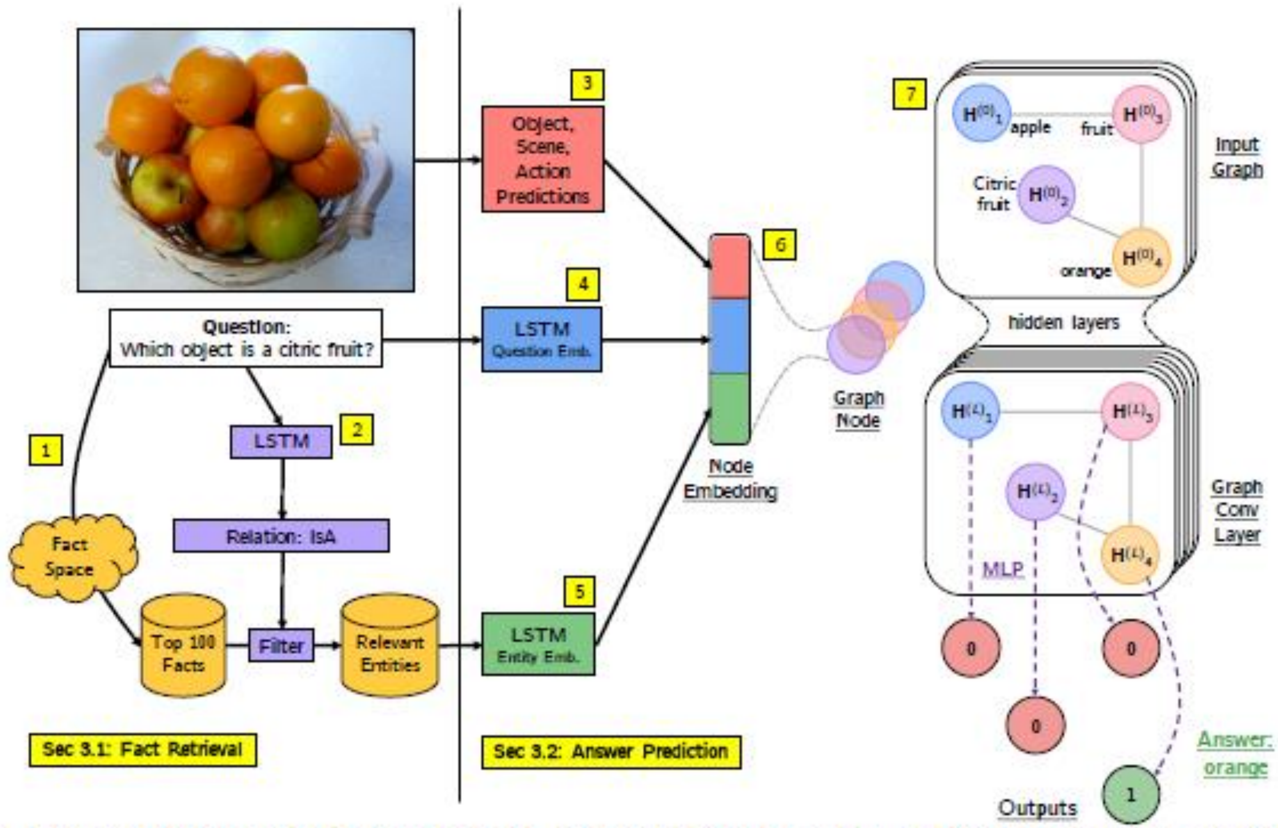
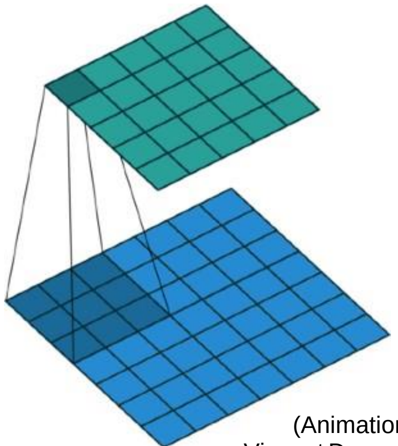


Figure 2: Outline of the proposed approach: Given an image and a question, we use a similarity scoring technique (1) to obtain relevant facts from the fact space. An LSTM (2) predicts the relation from the question to further reduce the set of relevant facts and its entities. An entity embedding is obtained by concatenating the visual concepts embedding of the image (3), the LSTM embedding of the question (4), and the LSTM embedding of the entity (5). Each entity forms a single node in the graph and the relations constitute the edges (6). A GCN followed by an MLP performs joint assessment (7) to predict the answer. Our approach is trained end-to-end.

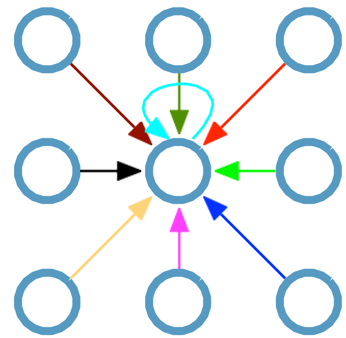
Narasimhan and Schwing, “Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering”, NeurIPS 2018

Graph convolutional networks

Recall: Single CNN layer with 3x3 filter:



(Animation by Vincent Dumoulin)



Update for a single pixel:

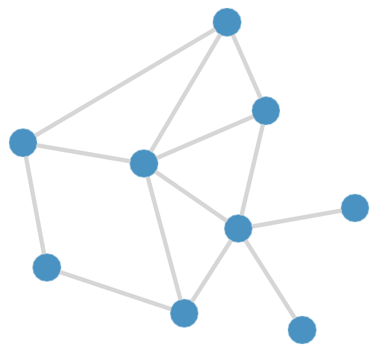
- Transform messages individually $\mathbf{W}_i \mathbf{h}_i$
- Add everything up $\sum_i \mathbf{W}_i \mathbf{h}_i$

Full update: $\mathbf{h}_4^{(l+1)} = \sigma\left(\mathbf{W}_0^{(l)}\mathbf{h}_0^{(l)} + \mathbf{W}_1^{(l)}\mathbf{h}_1^{(l)} + \cdots + \mathbf{W}_8^{(l)}\mathbf{h}_8^{(l)}\right)$

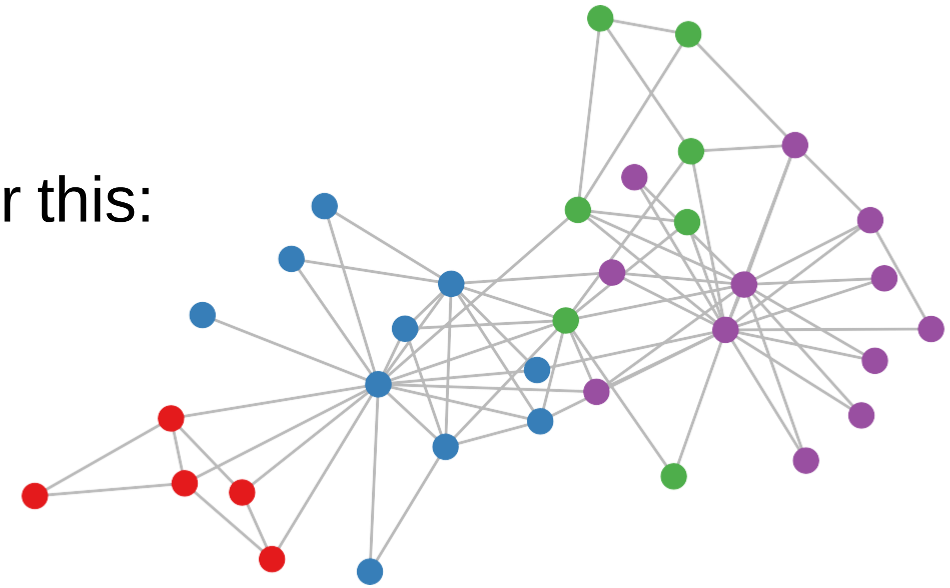
Kipf and Welling, "Semi-supervised learning with deep generative models", ICLR 2017 (slides by Thomas Kipf)

Graph convolutional networks

What if our data looks like this?



or this:



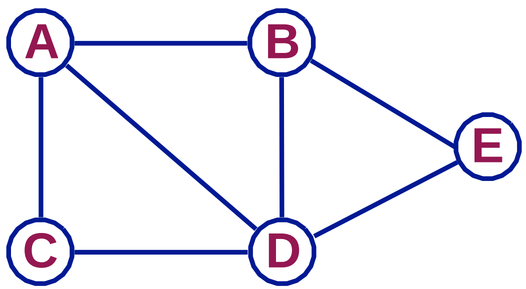
Real-world examples:

- Social networks
- World-wide-web
- Protein-interaction networks
- Telecommunication networks
- Knowledge graphs
- …

Kipf and Welling, "Semi-supervised learning with deep generative models", ICLR 2017 (slides by Thomas Kipf)

Graph convolutional networks

Graph: $G = (\mathcal{V}, \mathcal{E})$



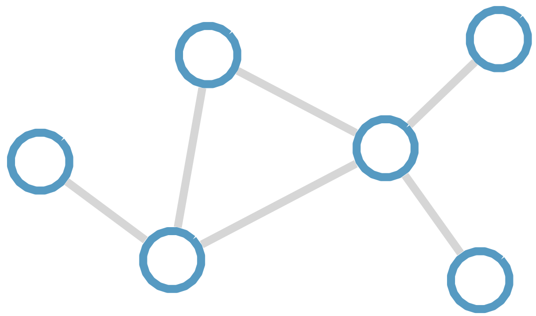
Adjacency matrix: A

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

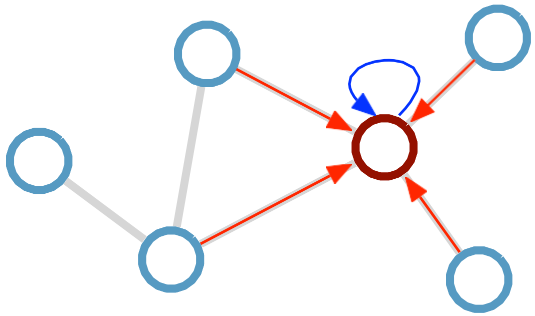
Kipf and Welling, “Semi-supervised learning with deep generative models”, ICLR 2017 (slides by Thomas Kipf)

Graph convolutional networks

Consider this undirected graph:



Calculate update for node in red:



Update rule: $$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

$\mathcal{N}_i$: neighbor indices

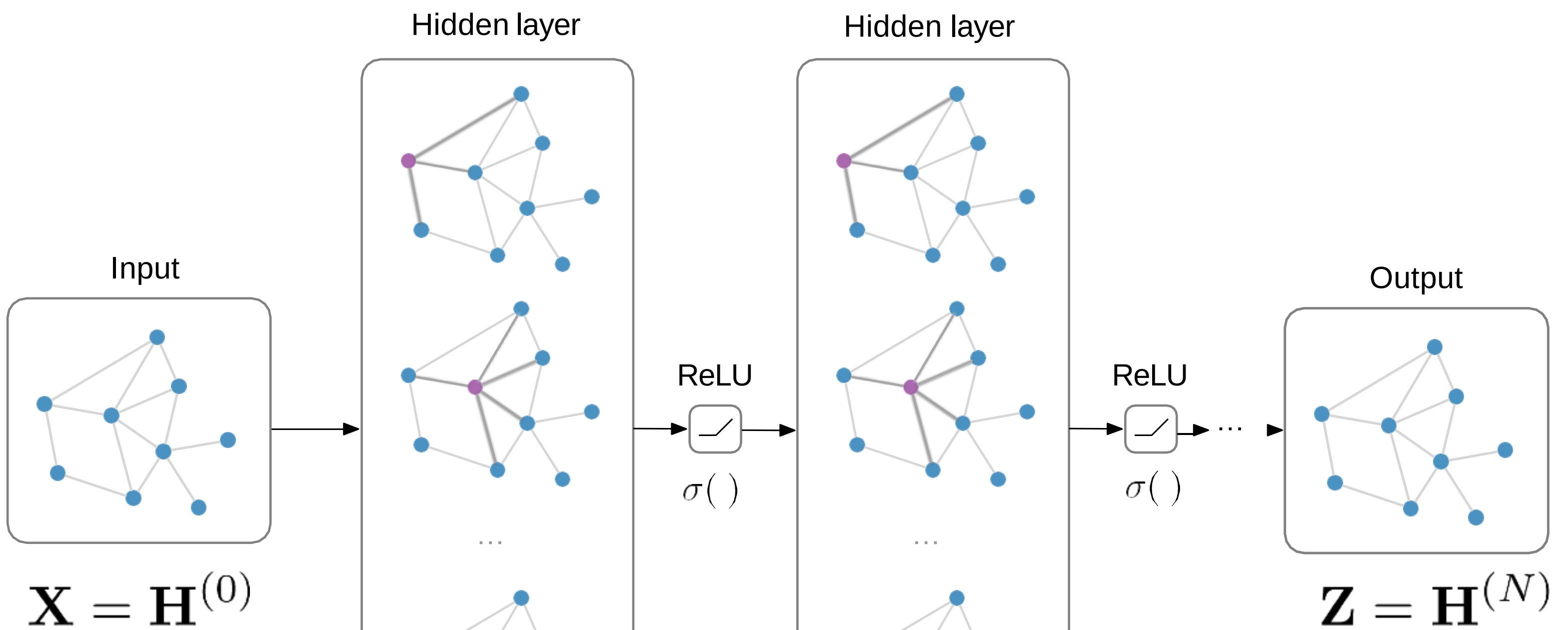
c_{ij} : norm. constant (per edge)

Note: We could also choose simpler or more general functions over the neighborhood

Kipf and Welling, "Semi-supervised learning with deep generative models", ICLR 2017 (slides by Thomas Kipf)

Graph convolutional networks

Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

Kipf and Welling, "Semi-supervised learning with deep generative models", ICLR 2017 (slides by Thomas Kipf)

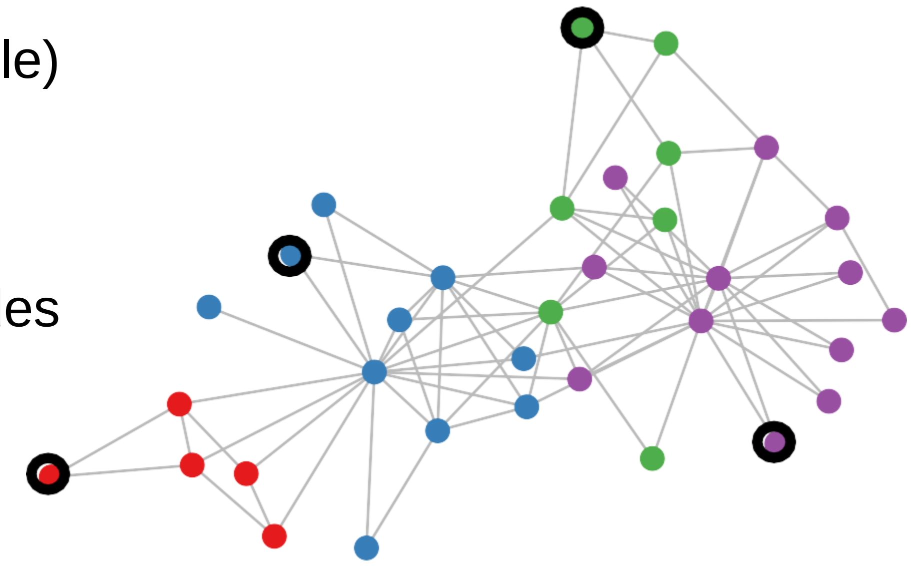
Graph convolutional networks

Semi-supervised classification on graphs

Setting:

Some nodes are labeled (black circle)
All other nodes are unlabeled

Task:
Predict node label of unlabeled nodes



Kipf and Welling, “Semi-supervised learning with deep generative models”, ICLR 2017 (slides by Thomas Kipf)

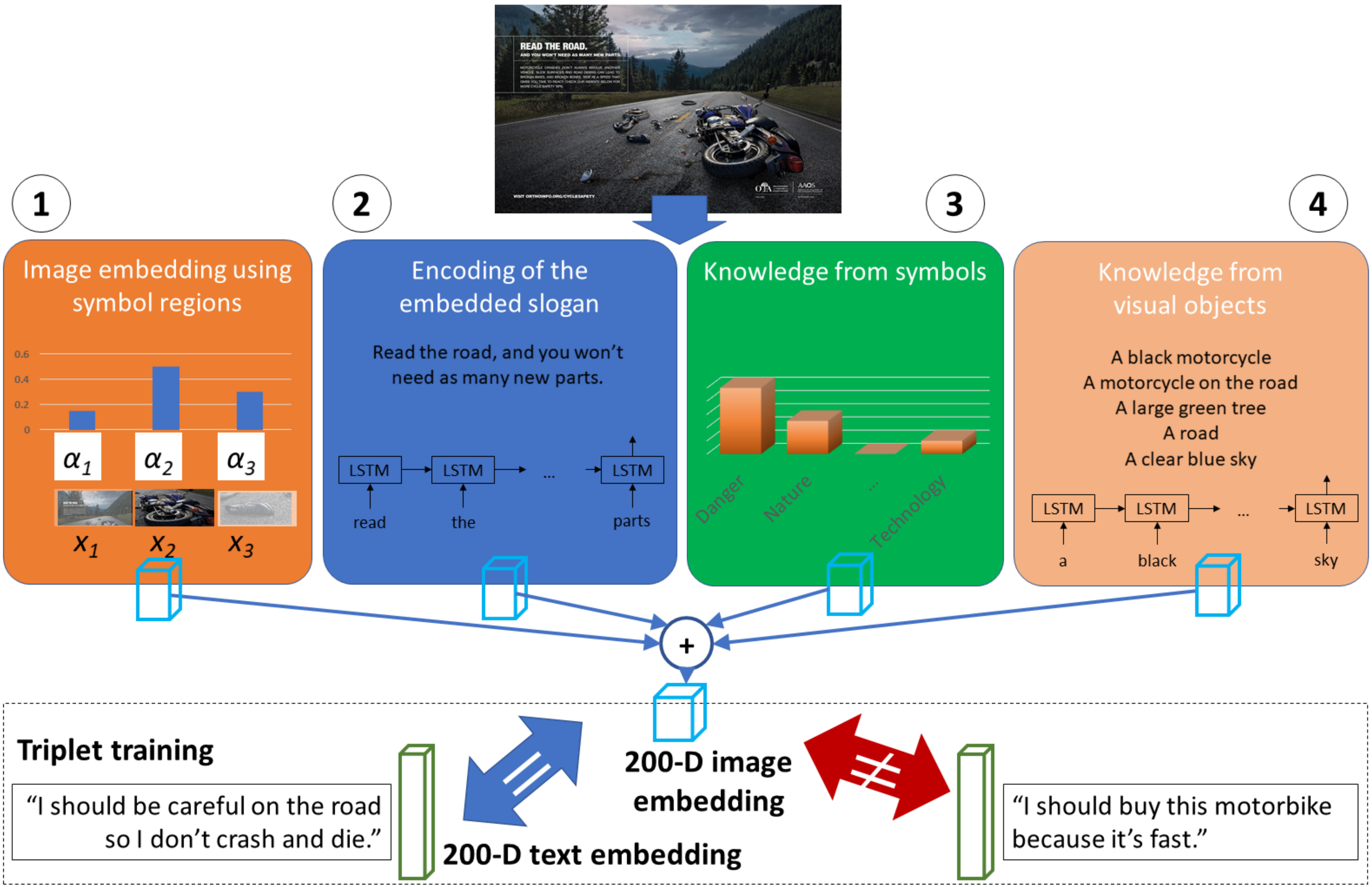
Decoding image advertisements

- What message does the ad convey (*action*), and what arguments does it provide for taking the suggested action (*reason*)?
- Multiple-choice task: Given *k* options for action-reason statements, pick one that matches the image

- I should drink evian because it helps you recover
- I should drink Evian because it will keep me like a baby
- I should buy Evian because it keeps us young

Hussain, Zhang, Zhang, Ye, Thomas, Agha, Ong and Kovashka, CVPR 2017

Retrieve the best action-reason statement



Ye et al., TPAMI 2019

Experimental results (image features only)

- We outperform prior art by a large margin, for both statement ranking and classification

	Rank (Lower ↓ is better)		Recall@3 (Higher ↑ is better)	
Method	PSA	Product	PSA	Product
2-way Nets	4.836 (± 0.090)	4.170 (± 0.023)	0.923 (± 0.016)	1.212 (± 0.004)
VSE	4.155 (± 0.091)	3.202 (± 0.019)	1.146 (± 0.017)	1.447 (± 0.004)
VSE++	4.139 (± 0.094)	3.110 (± 0.019)	1.197 (± 0.017)	1.510 (± 0.004)
Hussain-Ranking	*3.854* (± 0.088)	*3.093* (± 0.019)	*1.258* (± 0.017)	*1.515* (± 0.004)
ADVISE (ours)	**3.013** (± 0.075)	**2.469** (± 0.015)	**1.509** (± 0.017)	**1.725** (± 0.004)

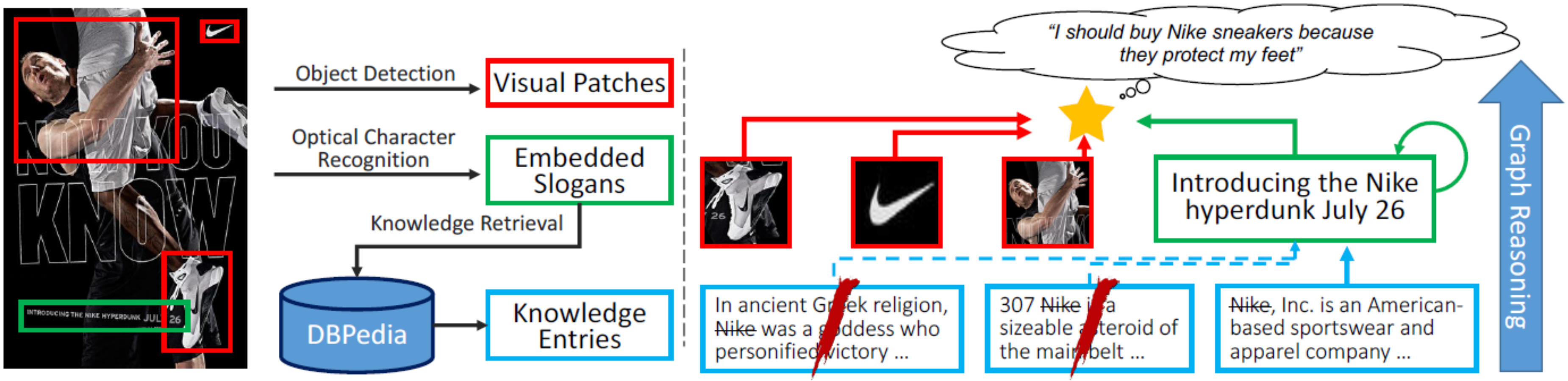
- Our methods accurately capture the rhetoric, even in deliberately confusing ads

VSE++ on Ads: I should wear Revlon makeup because it will make me more attractive"

ADVISE (ours): "I should stop smoking because it doesn't make me pretty"

Ye and Kovashka, ECCV 2018

Incorporating external knowledge



- Expand image representation using DBPedia
- Represent regions, slogans, KB nuggets in a graph
- Not all nuggets relevant
- All may be ignored due to non-generalizable shortcuts
- To prevent overfitting to shortcuts, we randomly mask parts of training samples (e.g. words in KB nugget, slogan)

Ye, Zhang and Kovashka, WACV 2021

Incorporating external knowledge

- Training via metric learning: match image to human-annotated action-reason statements
- Image representation is a graph
- Slogan node updates:

$$\mathbf{t}_i^{(1)} = \underbrace{\alpha_{i,0}\mathbf{t}_i^{(0)}}_{\text{original meaning}} + \underbrace{\sum_{j=1}^{|\phi(t_i)|} \alpha_{i,j}\mathbf{k}_{i,j}}_{\text{descriptions from extra knowledge}}$$

- Global node update:

$$\mathbf{h} = \underbrace{\sum_{i=1}^{|V|} \beta_i \mathbf{v}_i}_{\text{messages from proposals}} + \underbrace{\sum_{i=|V|+1}^{|V|+|T|} \beta_i \mathbf{t}_i^{(1)}}_{\text{messages from slogans}}$$

- Edge weights α, β allow model to choose what knowledge to use

Ye, Zhang and Kovashka, WACV 2021

Incorporating external knowledge

- We stochastically mask aspects of training data, to prevent model from relying too much on word-matching or object-matching
- Three strategies; can also learn how to mask:
  - M_t randomly drops a detected textual (T) slogan, with a probability of 0.5
  - M_s randomly sets the KB query words (e.g. "WWF" or "Nike") in the human-annotated statements (S) to the out-of-vocabulary token, with probability 0.5
  - M_k replaces the DBpedia queries in the retrieved knowledge contents with the out-of-vocabulary token

Ye, Zhang and Kovashka, WACV 2021

Incorporating external knowledge

- Outperform prior state of the art

Methods	Accuracy (%)
VSE [31]	62.0
ADNET [6]	65.0
ADVISE [31]	69.0
CYBERAGENT [18]	82.0
RHETORIC [32]	83.3
OURS	**87.3**

- Using external knowledge helps when data masked

Method	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10	Min Rank	Avg Rank	Med Rank
Results on the Challenge-15 task											
V,T	**87.3**	76.6	55.1	30.6	**28.4**	74.2	87.9	97.5	1.26	3.02	2.77
V,T+K	**87.3**	76.6	55.1	30.6	**28.4**	74.3	87.9	97.6	1.25	3.02	2.77
V,T+K(M_t,M_s,M_k)	**87.3**	**77.5**	**55.9**	**30.8**	**28.4**	**75.2**	**89.2**	**98.2**	**1.23**	**2.91**	**2.69**
Results on the Sampled-100 task											
V,T	79.8	66.5	46.9	26.2	26.0	64.4	74.9	83.5	2.38	7.52	5.86
V,T+K	80.0	67.0	47.0	26.1	26.0	64.9	75.1	83.4	2.29	7.49	5.81
V,T+K(M_t,M_s,M_k)	**80.2**	**67.9**	**47.9**	**26.8**	**26.1**	**65.8**	**76.6**	**85.4**	**2.14**	**6.56**	**5.19**
Results on the Sampled-500 task											
V,T	**65.5**	52.3	37.8	21.7	**21.3**	50.5	60.4	69.0	8.18	30.1	21.6
V,T+K	65.4	52.3	38.0	21.9	**21.3**	50.6	60.7	69.6	7.60	30.0	21.4
V,T+K(M_t,M_s,M_k)	64.8	**52.4**	**38.3**	**22.1**	21.1	**50.7**	**61.1**	**70.6**	**6.89**	**25.1**	**18.2**

Ye, Zhang and Kovashka, WACV 2021

Incorporating external knowledge

Quantitatively:
Without masking we retrieve relevant info with accuracy 25%, vs 54% with masking.

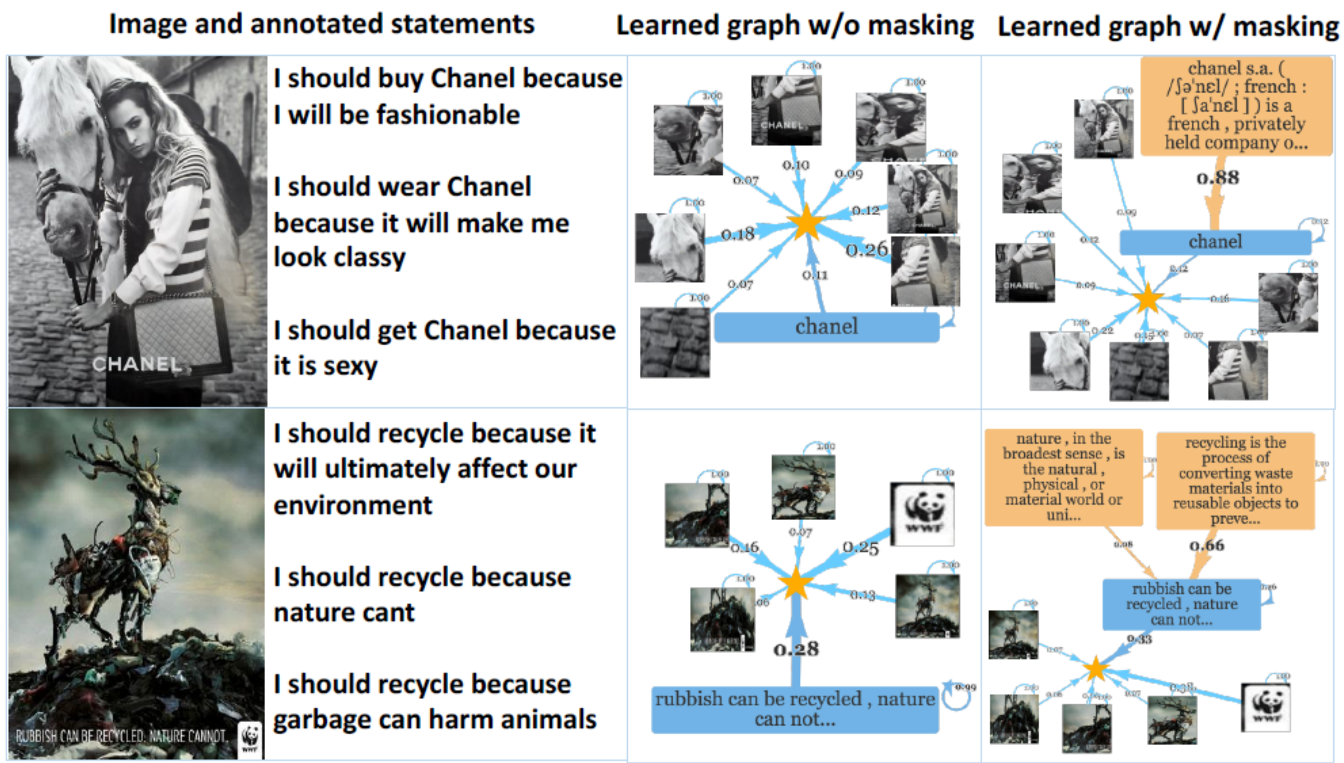


Fig. 4: **Examples of the learned graphs (best with zoom).** We show the ad image and annotated action-reason statements on the left, the graph learned without masking in the middle, and that learned with masking (our approach) on the right. We show slogans in blue, DBpedia comments in orange, and the global node as a star. **Arrow thickness is correlated with learned weights α, β.** For visualization we removed all edges with small weights (threshold=0.05). We see our method more effectively leverages external information.

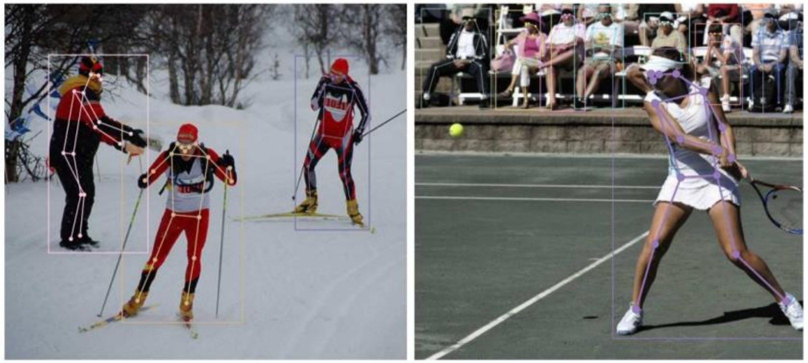
Ye, Zhang and Kovashka, WACV 2021

Plan for this lecture

- Learning the relation between images and text
  - Recurrent neural networks
  - Applications: Captioning
  - Transformers
- Reasoning: Visual question answering
  - Neuro-symbolic VQA
  - Graph convolutional networks
- Multimodal self-supervised learning

Multimodal self-supervised learning

Success of Supervised Learning



Pose estimation

[Towards Accurate Multi-person Pose Estimation in the Wild, Papandreou, Zhu, Kanazawa, Toshev, Tompson, Bregler and Murphy, CVPR17]

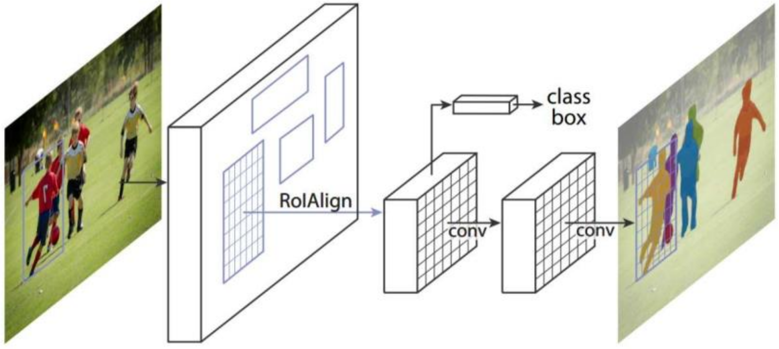
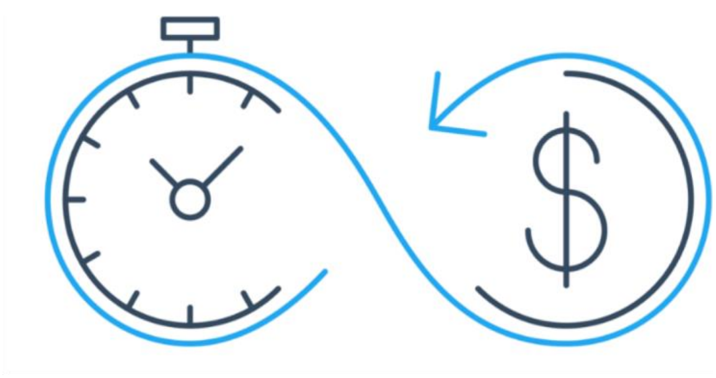


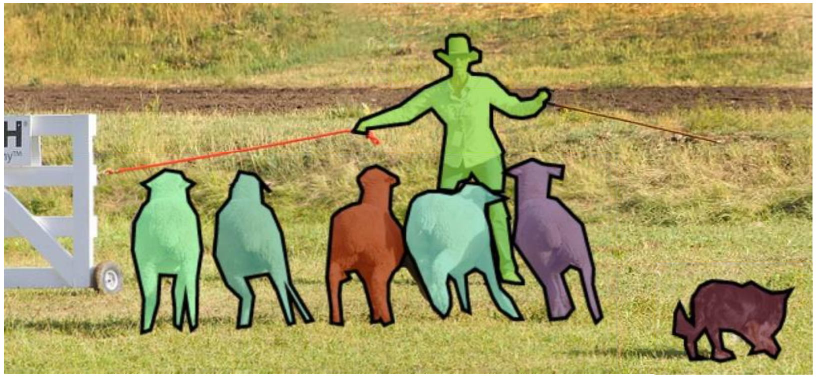
Image Segmentation

[Mask R-CNN, He, Gkioxari, Dollár, and Girshisck, ICCV17]

Issues of Supervised Learning

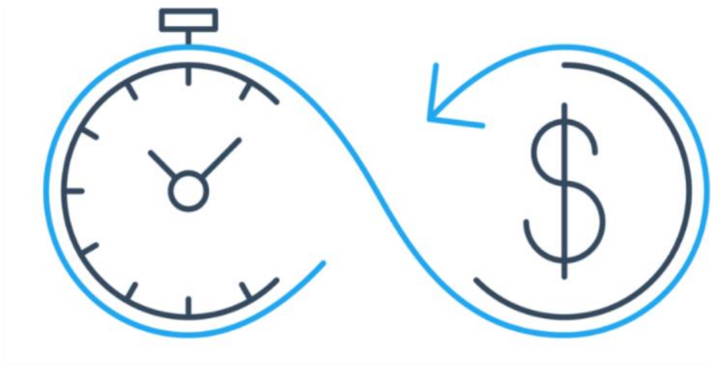


Labels are expensive



Agreement: definition? granularity?

Issues of Supervised Learning



Labels are expensive

Even more problematic for videos!

Weakly supervised learning
Use weaker and readily available source of supervision

Training info: **image level label**

[Barnard et al'03], [Joulin et al'10], [Deselaers et al'12], [Song et al'14], [Wang et al'14], [Cinbis et al'15], [Oquab et al'15], [Kantorov et al'16], [Bilen and Vedaldi'16]…

Can we use even weaker, cheaper supervision?

Adapted from Jean-Baptiste Alayrac

What are instructional videos?

- Depict complex, goal-oriented human activities (e.g. how to change a car tire)
- Multimodal: video and language
- Can be obtained at scale (e.g. on YouTube), without manual annotation

2

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, *ICCV19*





A. Miech* D. Zhukov* M. Tapaswi I. Laptev J. Sivic

*equal contribution

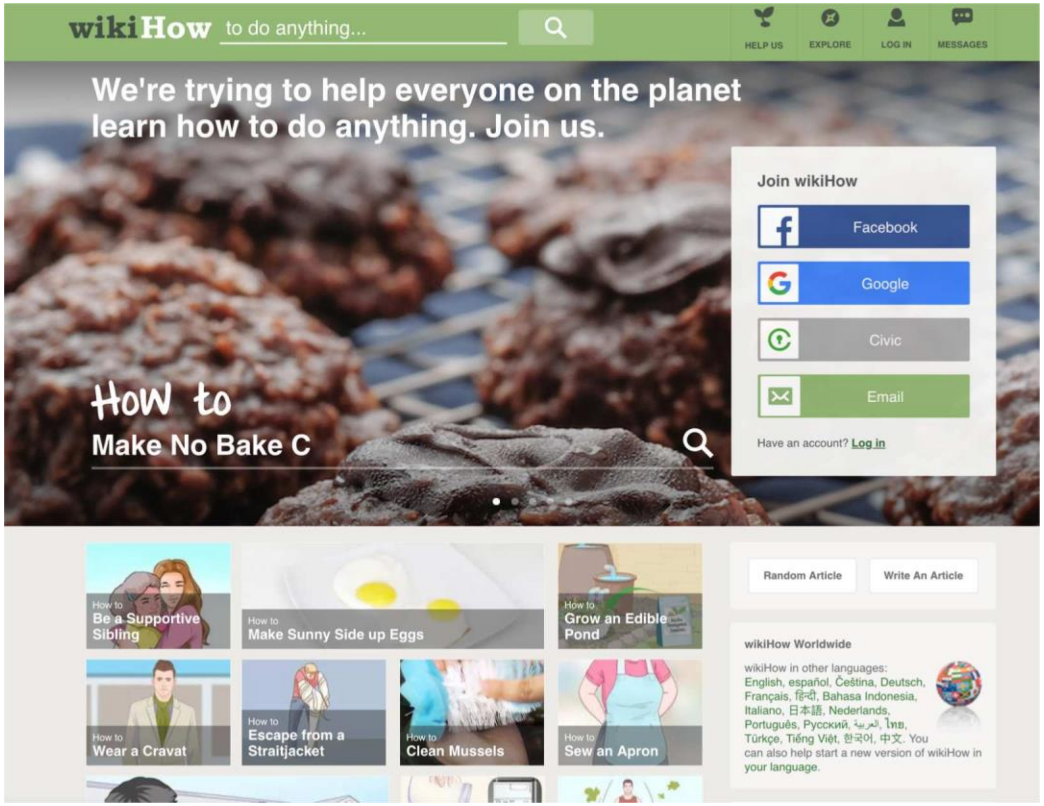
The HowTo100M dataset in numbers

- 23K human tasks scrapped from WikiHow
- 1.2M unique YouTube videos (duration 15 years)
- 136M clips with narration transcribed into text (mostly from ASR)
- Larger than any existing manually annotated captioning dataset

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [48]	10k	16k	10,000	82h	Home	2016
MSR-VTT [58]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [67]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [7]	40k	40k	432	55h	Home	2018
DiDeMo [15]	27k	41k	10,464	87h	Flickr	2017
M-VAD [52]	49k	56k	92	84h	Movies	2015
MPII-MD [43]	69k	68k	94	41h	Movies	2015
ANet Captions [26]	100k	100k	20,000	849h	Youtube	2017
TGIF [27]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [44]	128k	128k	200	150h	Movies	2017
How2 [45]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

How to collect HowTo100M?

Step 1 : WikiHow



Result: list of 130k tasks

...
How to be healthy
How to cook quinoa in a Rice Cooker
How to Sew an Apron
How to Break a Chain
How to April Fool your Girlfriend
...

Annotation cost: 0

Jean-Baptiste Alayrac

How to collect HowTo100M?

Step 2 : Filter task by verb to keep visual tasks

Result: list of 23k tasks

...

~~How to Be healthy~~

✓ How to **Cook** quinoa in a Rice Cooker

✓ How to **Sew** an Apron

✓ How to **Break** a Chain

~~How to April Fool your Girlfriend~~

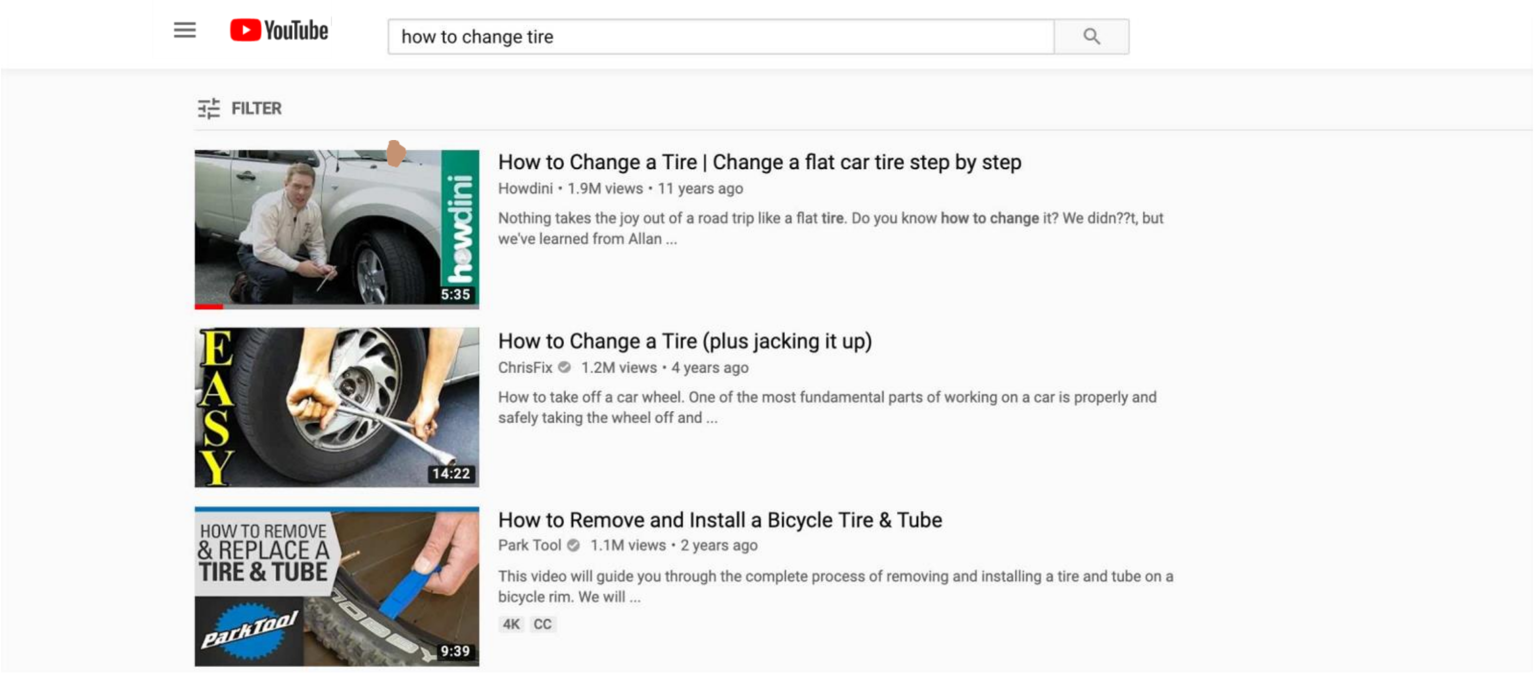
...

Annotation cost: 8 hours for Antoine

How to collect HowTo100M?

Step 3 : YouTube queries for videos with captions

Result: 1.2M unique videos

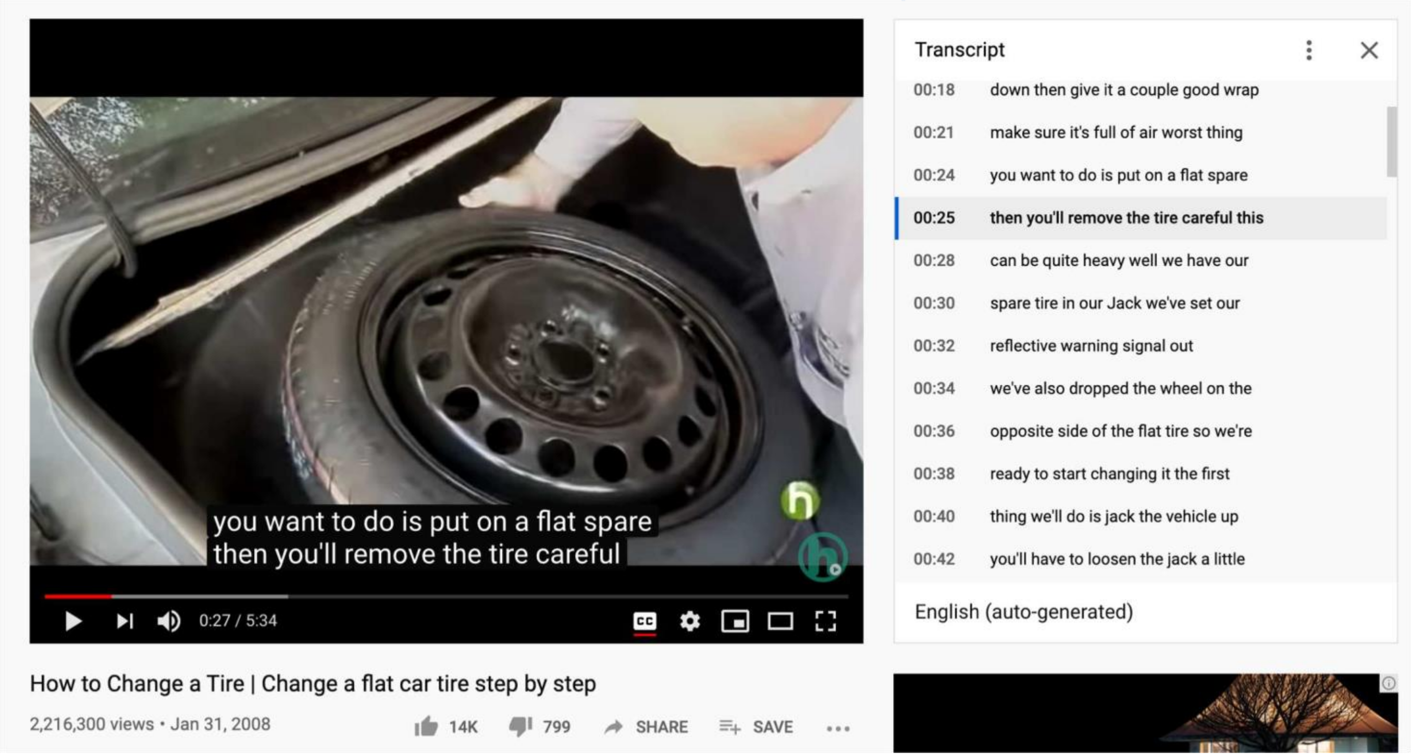


Annotation cost: 0

How to collect HowTo100M?

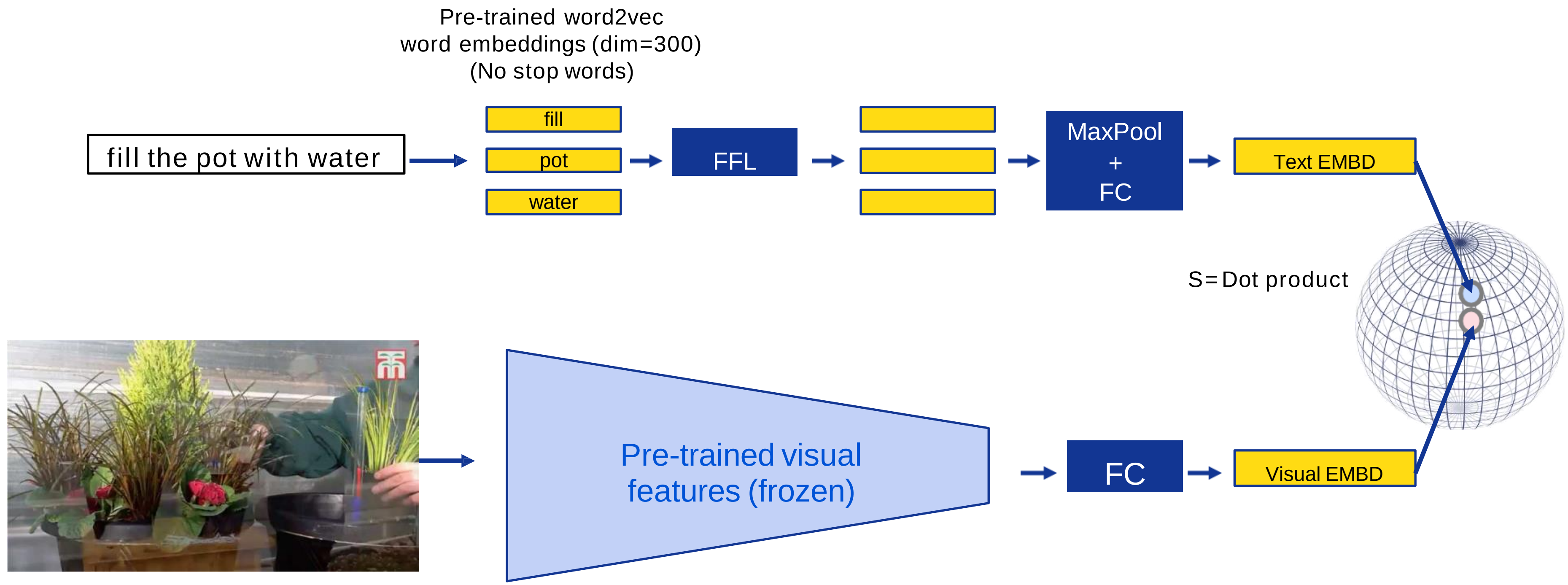
Step 4 : Create clips

Result: 136M narrated clips



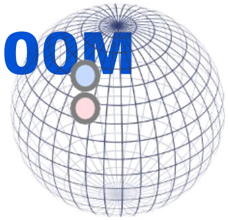
Annotation cost: 0

Learning a visual-text embedding on HowTo100M



DeViSE: A Deep Visual-Semantic Embedding Model, Frome et al. NeurIPS2013

Learning a visual-text embedding on HowTo100M

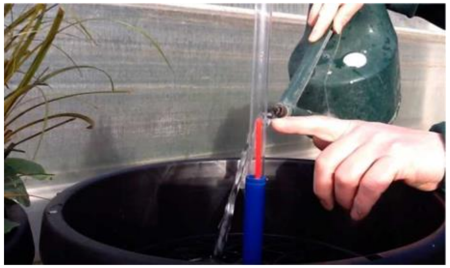


$$S_{i,j} = S(X_i, Y_j)$$ (dot product)

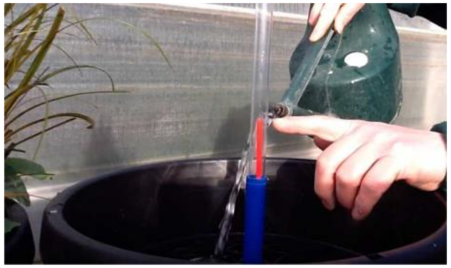
$$\forall (i,j),\ \ j \neq i, S_{i,i} > S_{i,j}\,, S_{i,i} > S_{j,i}$$

$$L = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}\Big[\max(0, m+S_{i,j}-S_{i,i})+\max(0, m+S_{j,i}-S_{i,i})\Big]$$

…fill pot water …



…fill pot water …



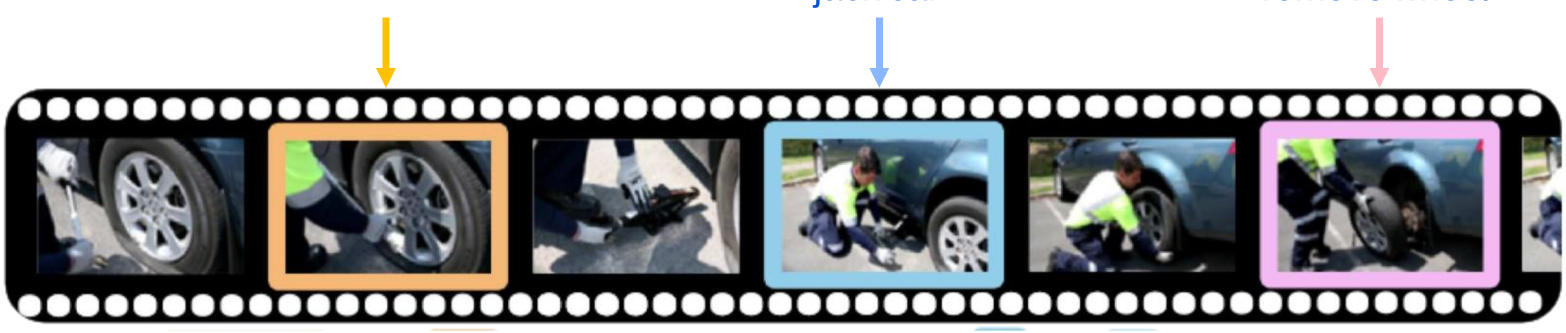
…these nice plants…

Evaluation procedure

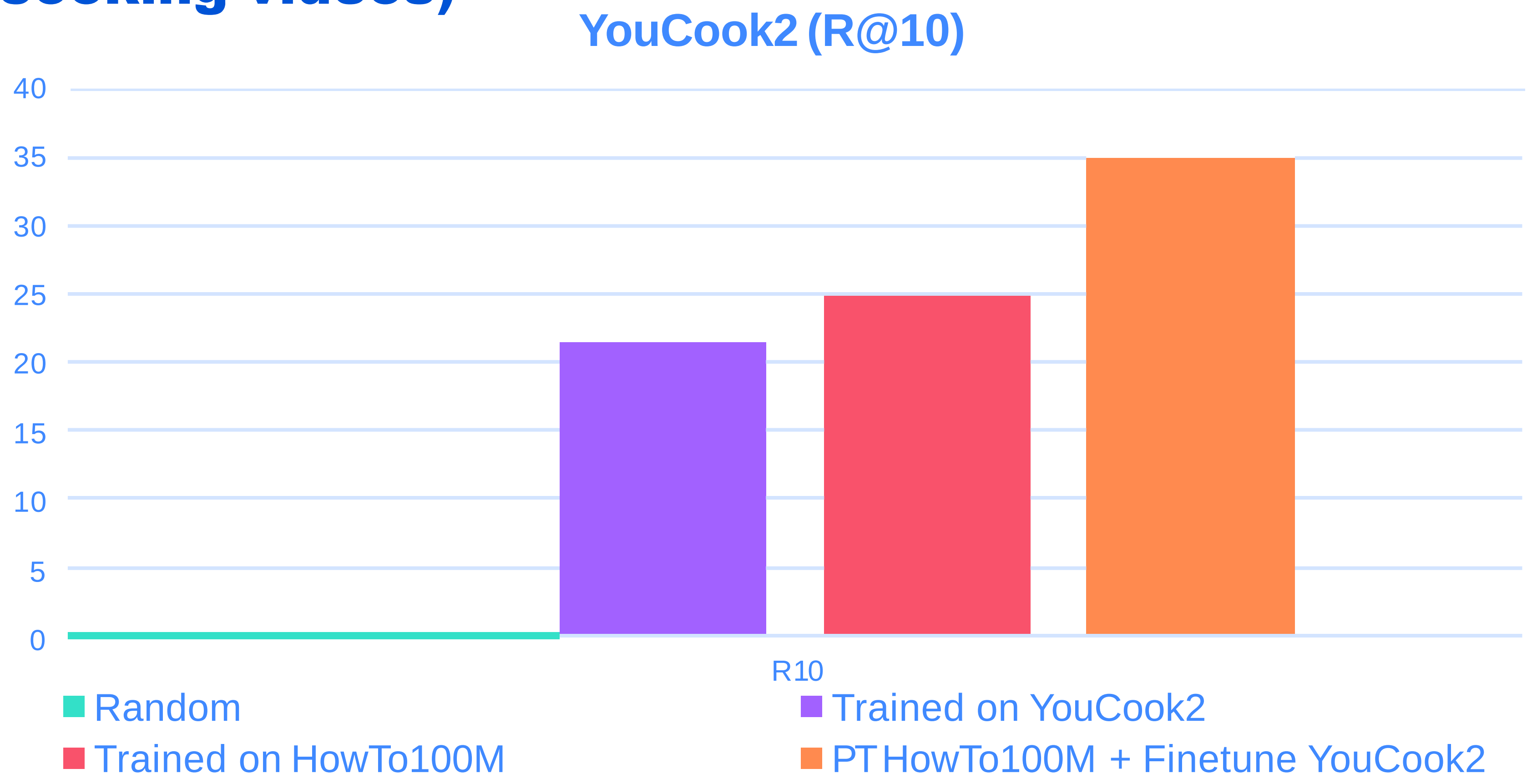
- Text to video retrieval: YouCook2, MSRVTT, LSMDC

Answering the phone

- Action localization: CrossTask



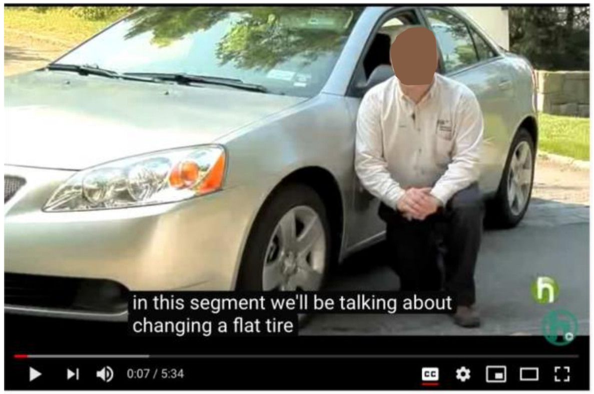
Within domain: YouCook2 retrieval (YouTube cooking videos)

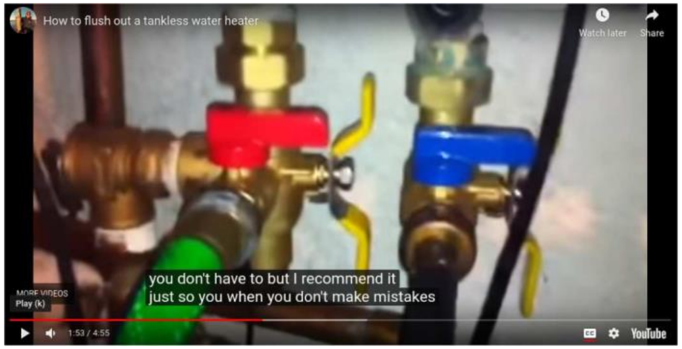


Jean-Baptiste Alayrac

Future directions

- Dealing with the noise. In 50% of the cases, video and narration are not matching. Something should be done!





- Still relying on pretrained features (obtained from Kinetics or ImageNet) so the story is not complete.

  The dream: end to end learning directly from HowTo100M.

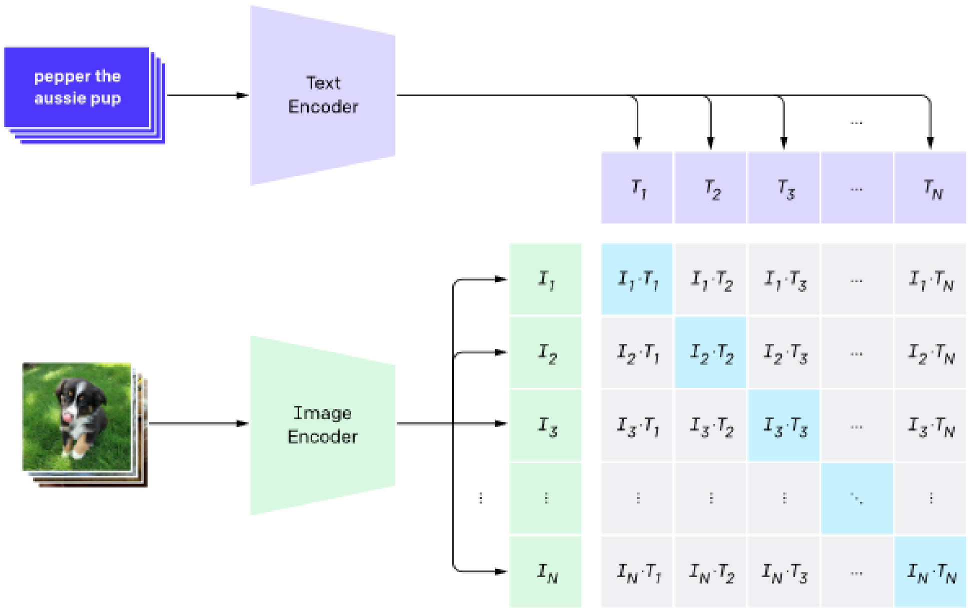
DALLE: Generating Images from Text Description

Capability: combining unrelated concepts in plausible ways

https://openai.com/blog/dall-e/, slides by Meiqi Guo

CLIP: Contrastive Language-Image Pre-training

- Given a batch of N (image, text) pairs, CLIP is trained to predict which of the N*N possible pairings actually occurred.
- Contrastive representation learning: more compitationally efficient

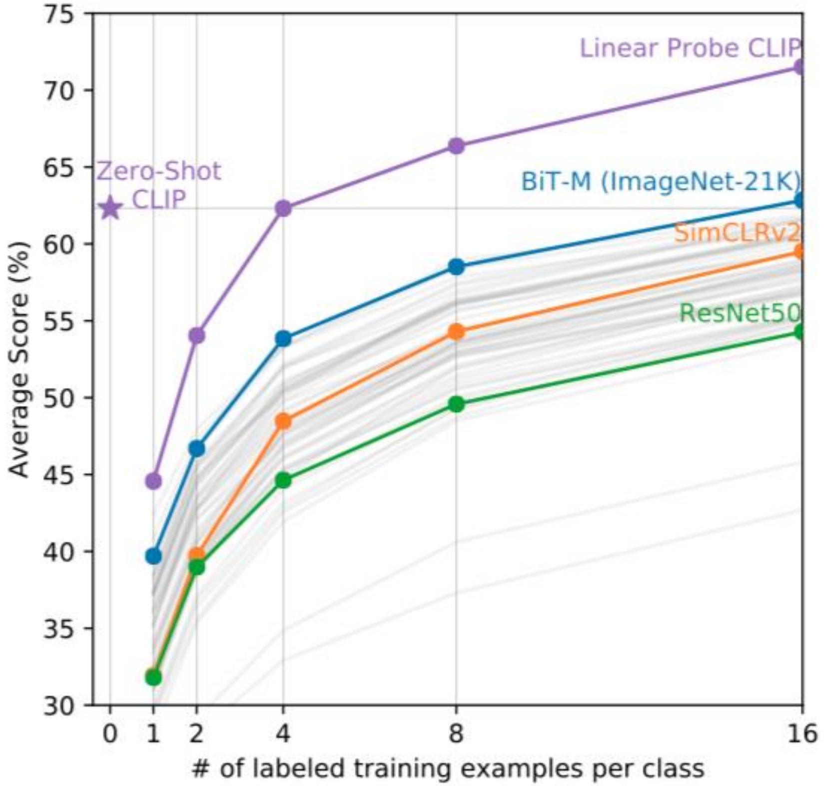



```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

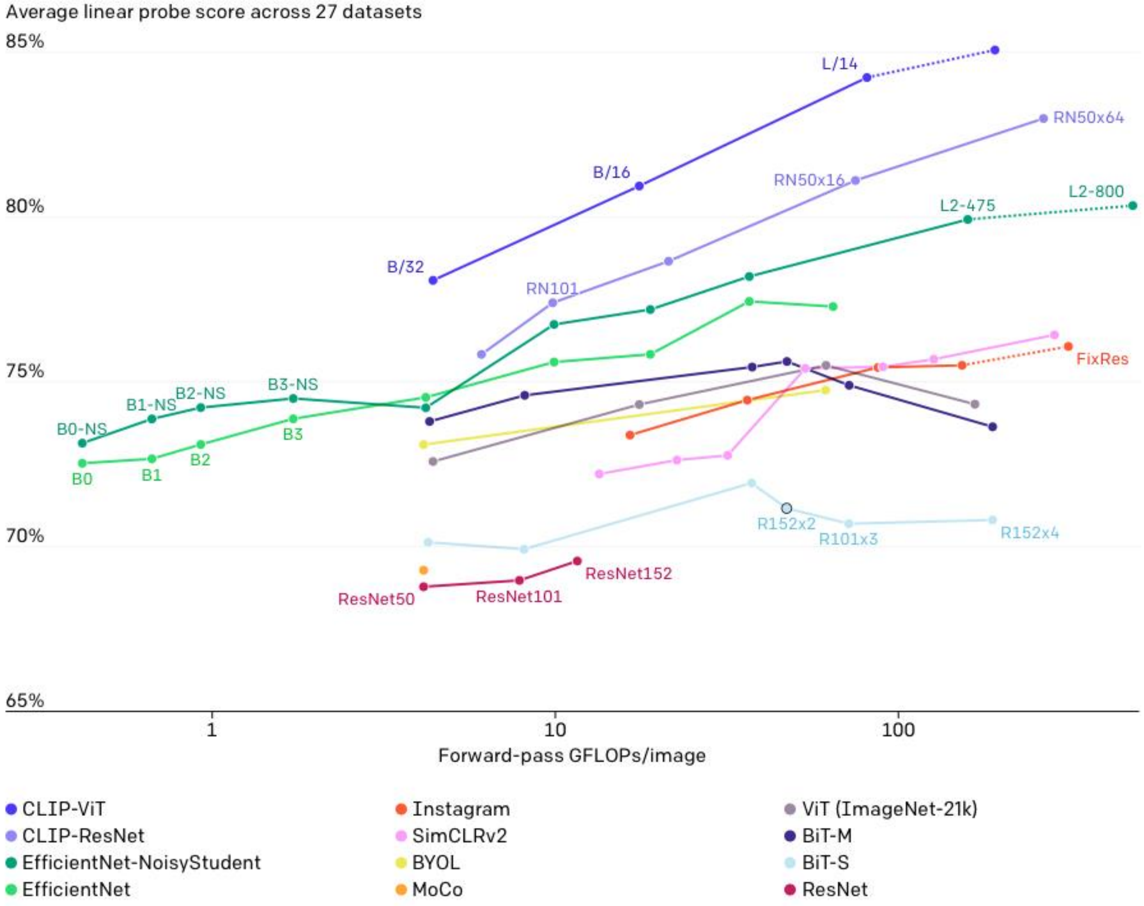
https://openai.com/blog/clip/, slides by Meiqi Guo

Results on Few-shot Classification



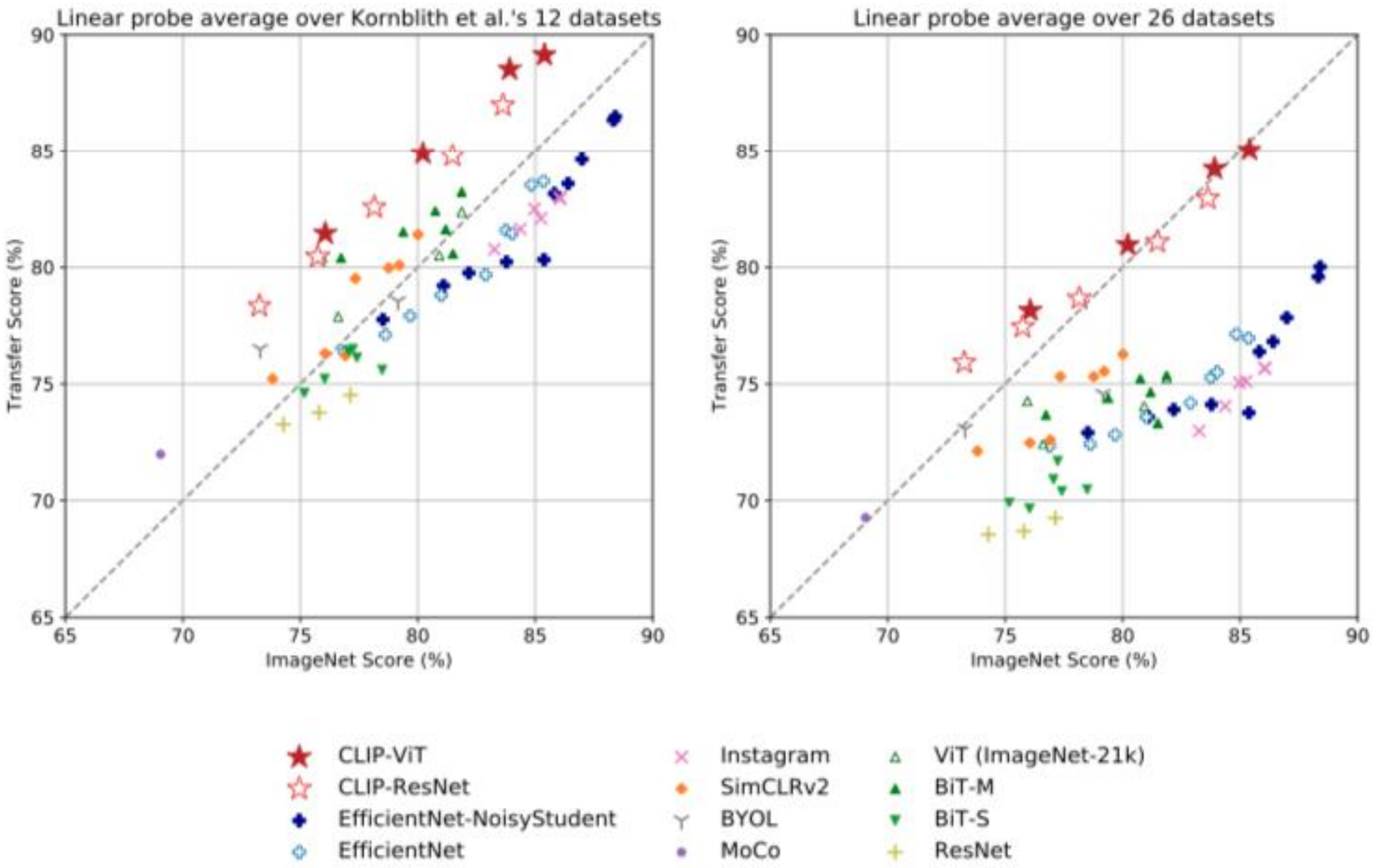
https://openai.com/blog/clip/, slides by Meiqi Guo

Results on Representation Learning



https://openai.com/blog/clip/, slides by Meiqi Guo

Results on Distribution Shifting



https://openai.com/blog/clip/, slides by Meiqi Guo