Language, Vision and Sequences

EECS 442 – David Fouhey

Winter 2023, University of Michigan

http://web.eecs.umich.edu/~fouhey/teaching/EECS442_W23/

Quick – what's this?

Dog image credit: T. Gupta

Previously on EECS 442

Feature vector from image

						x_i		Wx_i	
Cat weight vector	0.2	-0.5	0.1	2.0	1.1	56		-96.8	Cat score
Dog weight vector	1.5	1.3	2.1	0.0	3.2	231	→	437.9	Dog score
Hat weight vector	0.0	0.3	0.2	-0.3	-1.2	24		61.95	Hat score
						2			
						1			

W

Wx_i

Weight matrix a collection of scoring functions, one per class

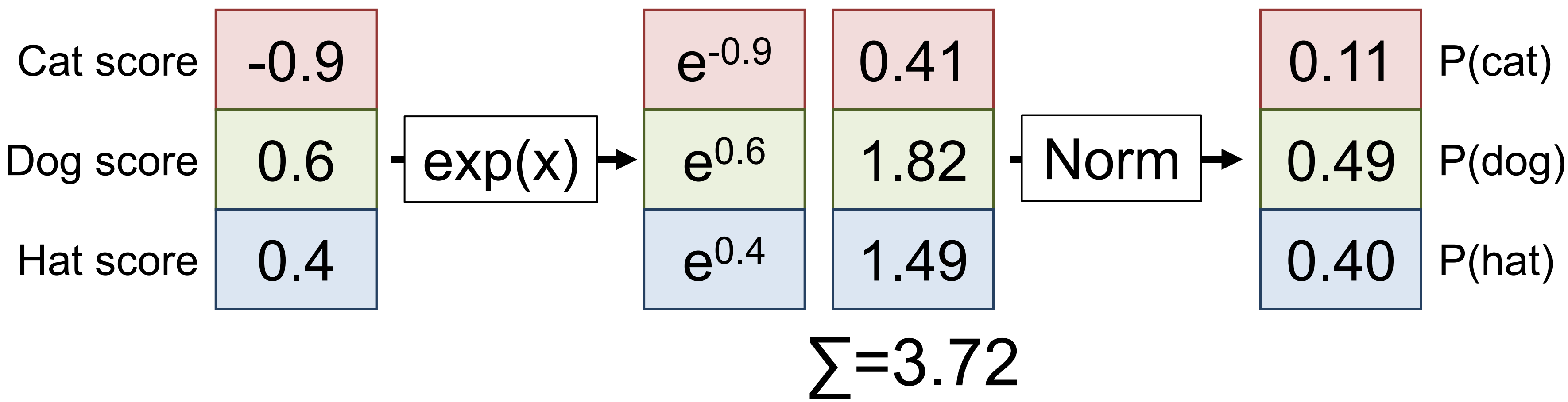
x_i

Prediction is vector where jth component is “score” for jth class.

Diagram by: Karpathy, Fei-Fei

Previously on EECS 442

Converting Scores to "Probability Distribution"



Generally P(class j): $\dfrac{\exp((Wx)_j)}{\sum_k \exp((Wx)_k)}$

What’s a Big Issue?

Is it a dog? Is it a hat?

Take 2

Feature vector from image

$$
\begin{array}{r}
\text{Cat weight vector} \\
\text{Dog weight vector} \\
\text{Hat weight vector}
\end{array}
\underbrace{\begin{bmatrix} 0.2 & -0.5 & 0.1 & 2.0 & 1.1 \\ 1.5 & 1.3 & 2.1 & 0.0 & 3.2 \\ 0.0 & 0.3 & 0.2 & -0.3 & -1.2 \end{bmatrix}}_{\boldsymbol{W}}
\underbrace{\begin{bmatrix} 56 \\ 231 \\ 24 \\ 2 \\ 1 \end{bmatrix}}_{\boldsymbol{x_i}}
\rightarrow
\underbrace{\begin{bmatrix} -96.8 \\ 437.9 \\ 61.95 \end{bmatrix}}_{\boldsymbol{Wx_i}}
\begin{array}{l}
\text{Cat score} \\
\text{Dog score} \\
\text{Hat score}
\end{array}
$$

Weight matrix a collection of scoring functions, one per class

Prediction is vector where jth component is “score” for jth class.

Diagram by: Karpathy, Fei-Fei

Take 2

Converting Scores to “Probability Distribution”

Cat score	-1.9		0.13	P(cat)
Dog score	1.2	sgm(x) →	0.77	P(dog)
Hat score	0.9		0.71	P(hat)

77% dog
71% hat
13% cat?

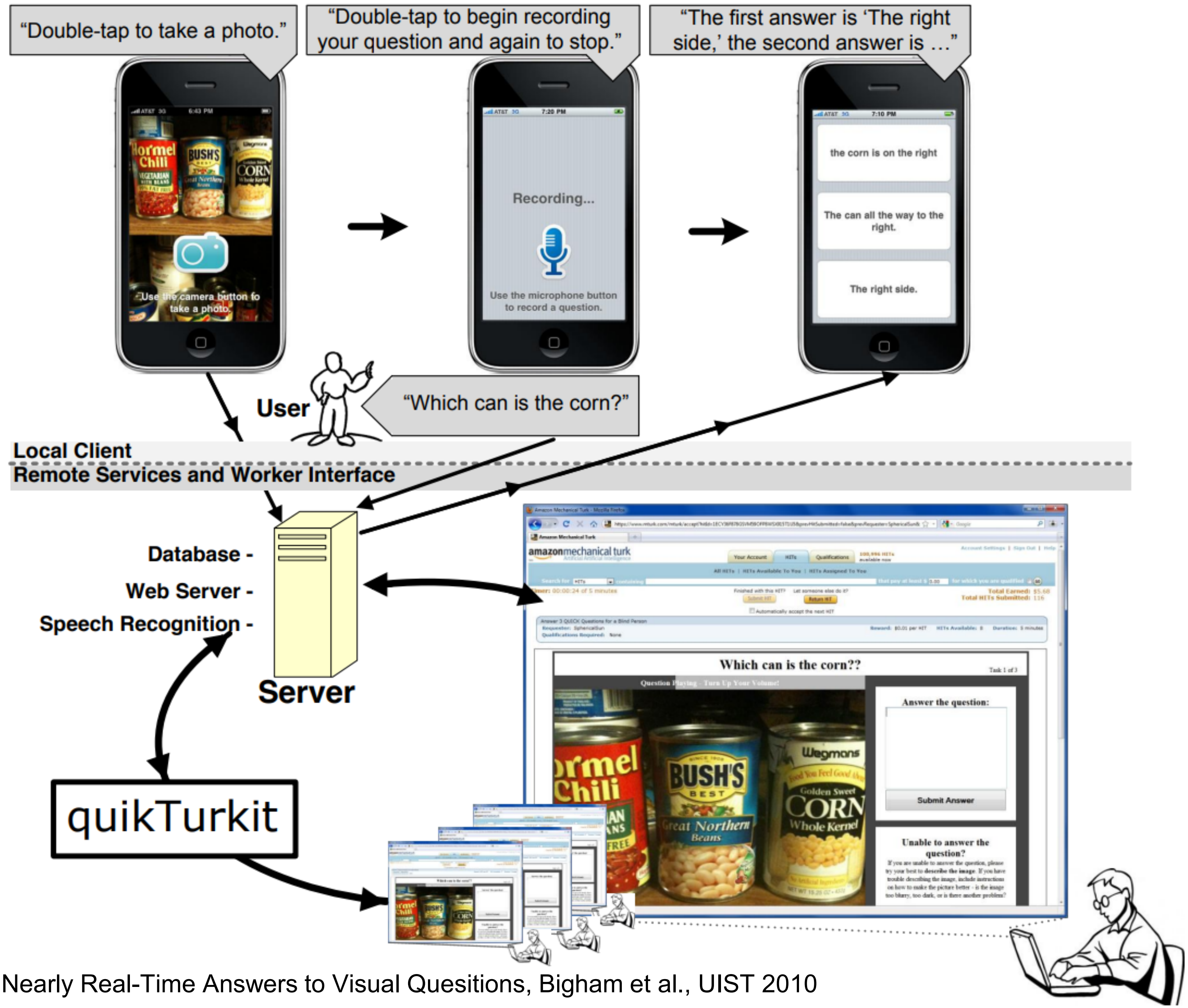
Hmm…

- We’d like to say: “dog with a hat” or “husky wearing a hat” or something else.
- Naïve approach (given N words to choose from and up to C words). **How many?**
- $\sum_{i=1}^{C} N^i$ classes to choose from ($\sim N^i$)
- N=10k, C=5 -> 100 billion billion
- Can’t train 100 billion billion classifiers

Hmm…

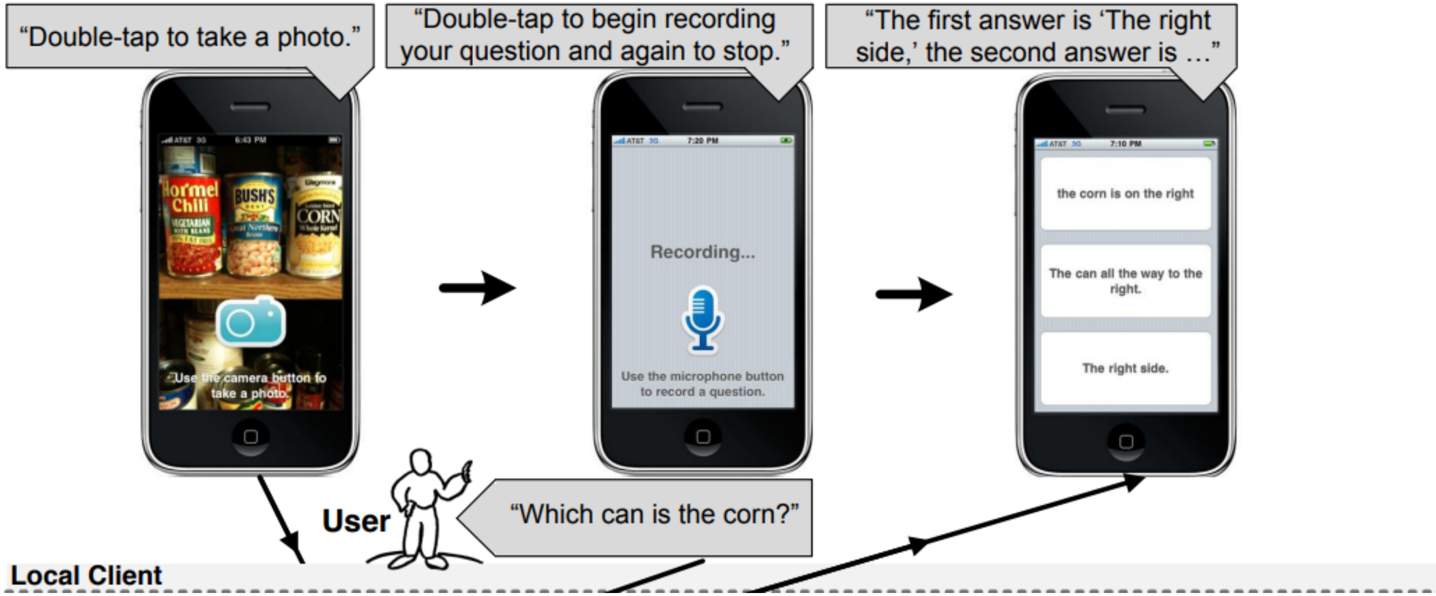
- Pick N-word dictionary, call them class 1, …, N
- New goal: emit sequence of C N-way classification outputs
- Dictionary could be:
  - All the words that appear in training set
  - All the ascii characters
  - Typically includes special “words”: START, END, UNK

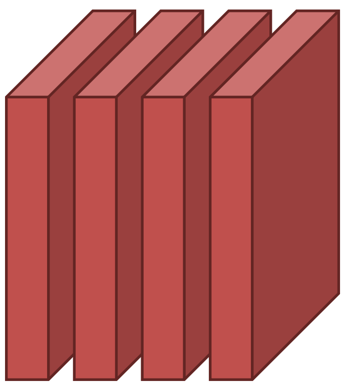
VizWiz



VizWiz: Nearly Real-Time Answers to Visual Quesitions, Bigham et al., UIST 2010

VizWiz

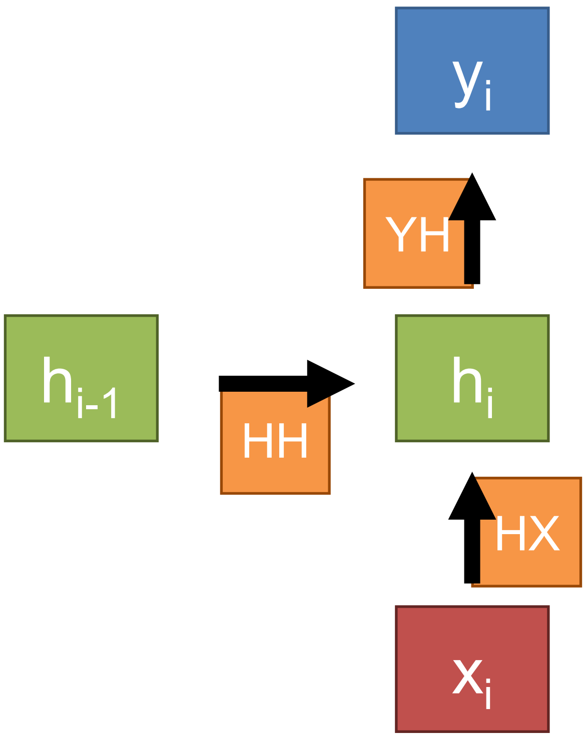




Deep learning system

Really interesting area. See Captioning Images Taken by People Who Are Blind, Gurari et al. ECCV 2020 and vizwiz.org

Option 1 – Sequence Modeling



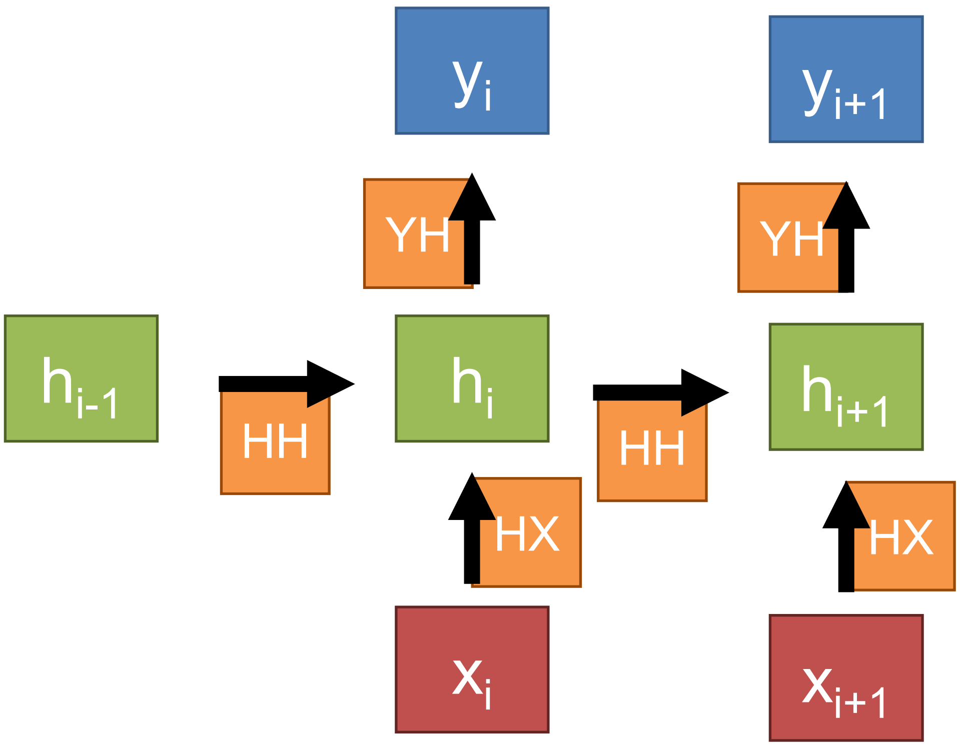
Output at i is linear transformation of hidden state

$$\boldsymbol{y_i = W_{yh}\, h_i}$$

Hidden state at i is linear function of previous hidden state and input at i, + nonlinearity

$$\boldsymbol{h_i = \sigma(W_{hx} x_i + W_{hh} h_{\mathbf{i-1}})}$$

Option 1 – Sequence Modeling

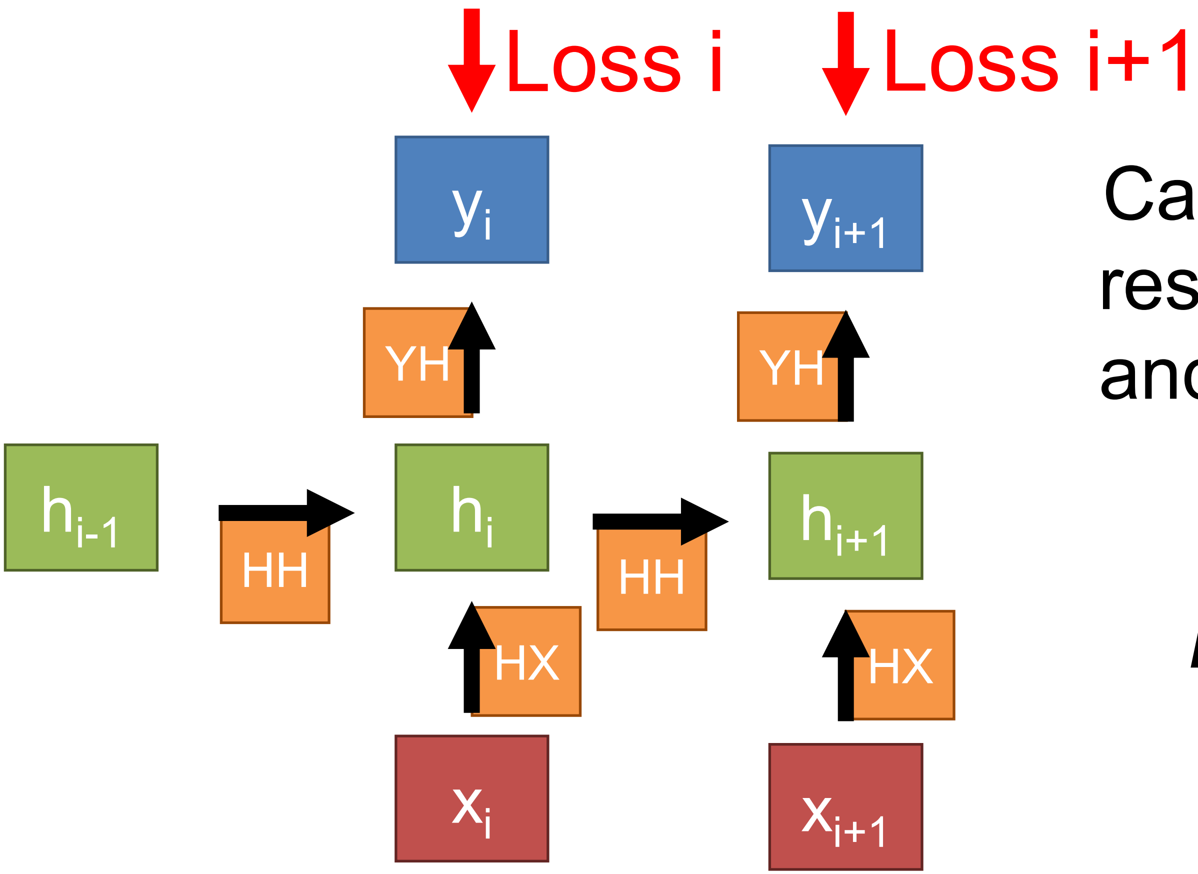


Can stack arbitrarily to create a function of multiple inputs with multiple outputs that's in terms of parameters W_{HX}, W_{HH}, W_{YH}

$$\boldsymbol{y_i = W_{yh}\, h_i}$$

$$\boldsymbol{h_i = \sigma(W_{hx} x_i + W_{hh} h_{i-1})}$$

Option 1 – Sequence Modeling



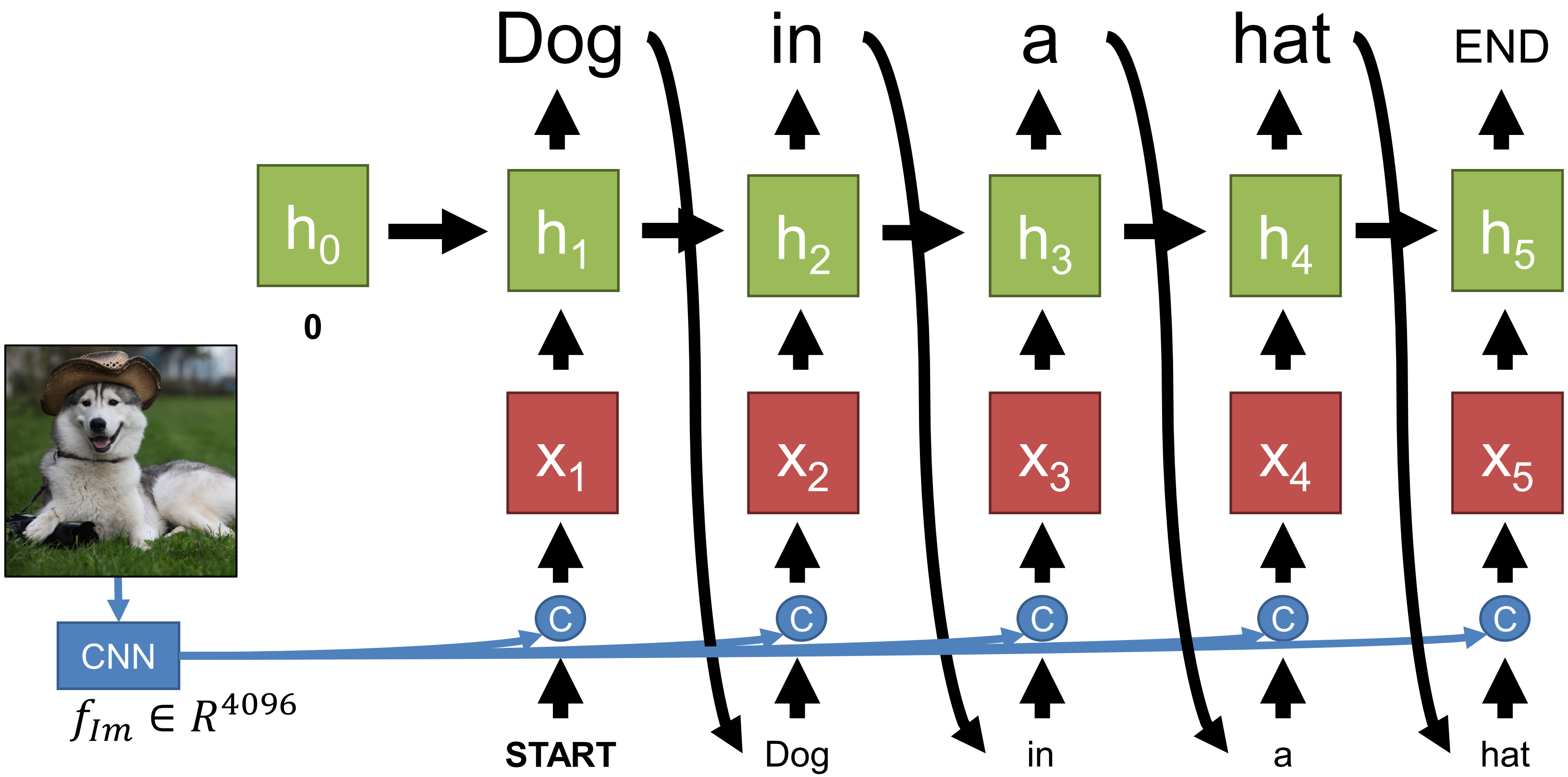
Can define a loss with respect to each output and differentiate wrt to all the weights

Backpropagation through time

$$\boldsymbol{y_i = W_{yh}\, h_i}$$

$$\boldsymbol{h_i = \sigma(W_{hx}x_i + W_{hh}h_{i-1})}$$

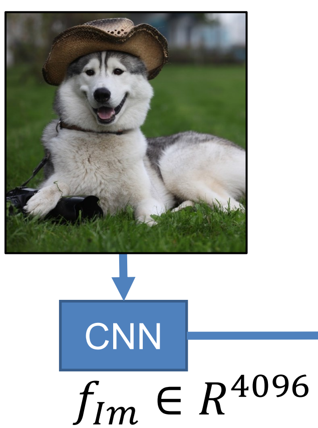
Captioning

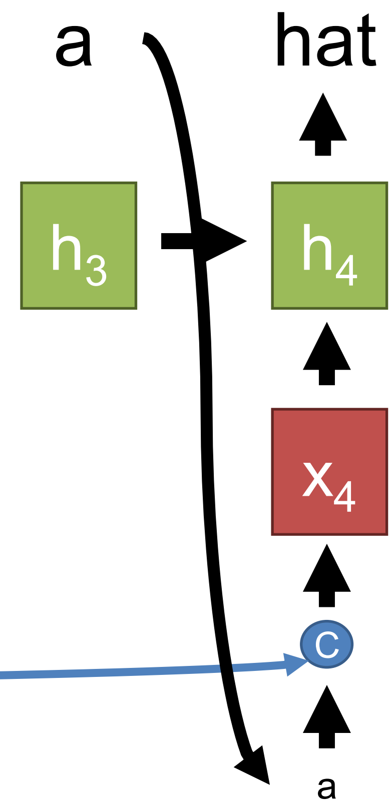


Dog
in
a
hat
END
h0
h1
h2
h3
h4
h5
0
x1
x2
x3
x4
x5
C
CNN
$f_{Im} \in R^{4096}$
START
Dog
in
a
hat

Captioning

Each step: look at input and hidden state (more on that in a second) and decide output. Can learn through CNN!





Results

A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surfboard.

A baseball game in progress with the batter up to plate.

A brown bear standing on top of a lush green field.

A person holding a cell phone in their hand.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Donahue et al. TPAMI, CVPR 2015.

Results

A close up of a person brushing his teeth.

A woman laying on a bed in a bedroom.

A black and white cat is sitting on a chair.

A large clock mounted to the side of a building.

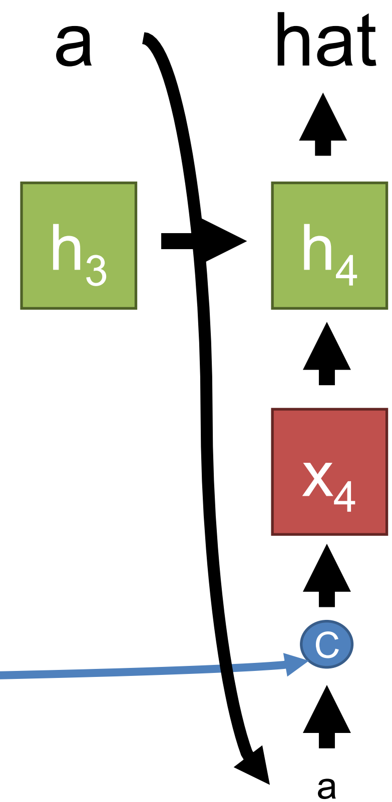
A bunch of fruit that are sitting on a table.

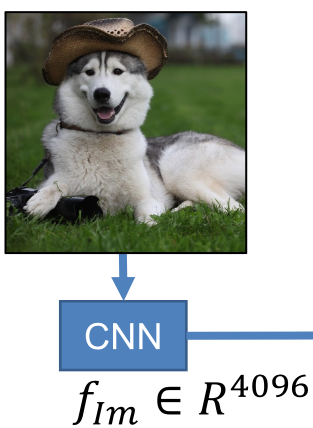
A toothbrush holder sitting on top of a white sink.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Donahue et al. TPAMI, CVPR 2015.

Captioning – Looking at Each Step

Why might this be better than doing billions of classification problems?





What Goes On Inside?

- Great repo for playing with RNNs (Char-RNN)
- https://github.com/karpathy/char-rnn
- (Or search char-rnn numpy)
- Tokens are just the characters that appear in the training set

Result credit: A. Karpathy

Sample Trained on Linux Code

```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

Result credit: A. Karpathy

Sample Trained on Names

Rudi Levette Berice Lussa Hany Mareanne Chrestina Carissy Marylen Hammine Janye Marlise Jacacrie Hendred Romand Charienna Nenotto Ette Dorane Wallen Marly Darine Salina Elvyn Ersia Maralena Minoria Ellia Charmin Antley Nerille Chelon Walmor Evena Jeryly Stachon Charisa Allisa Anatha Cathanie Geetra Alexie Jerin Cassen Herbett Cossie Velen Daurenge Robester Shermond Terisa Licia Roselen Ferine Jayn Lusine Charyanne Sales

Result credit: A. Karpathy

What Goes on Inside

Outputs of an RNN. Blue to red show timesteps where a given cell is active. **What’s this?**

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
          struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
         (void **)&df->lsm_rule);
  /* Keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \'%s\' is invalid\n",
      df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

Result credit: A. Karpathy

What Goes on Inside

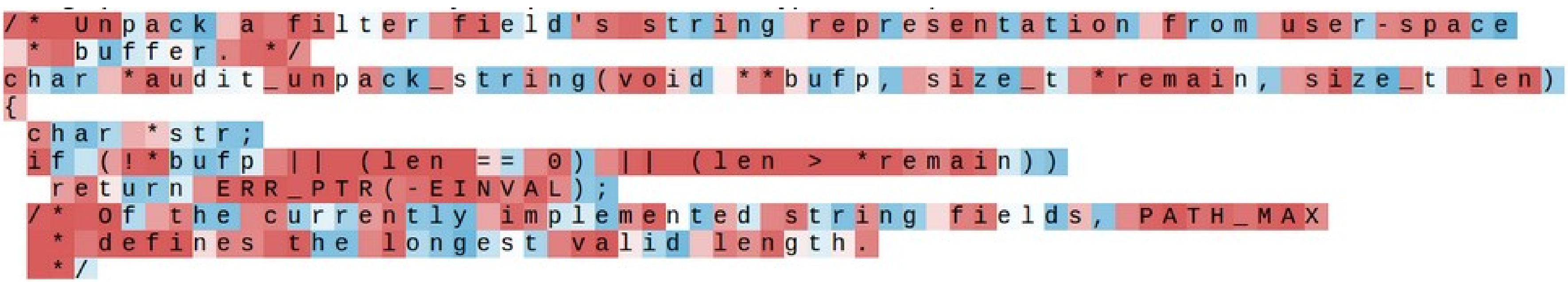
Outputs of an RNN. Blue to red show timesteps where a given cell is active. **What's this?**

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

Result credit: A. Karpathy

What Goes on Inside

Outputs of an RNN. Blue to red show timesteps where a given cell is active. **What's this?**



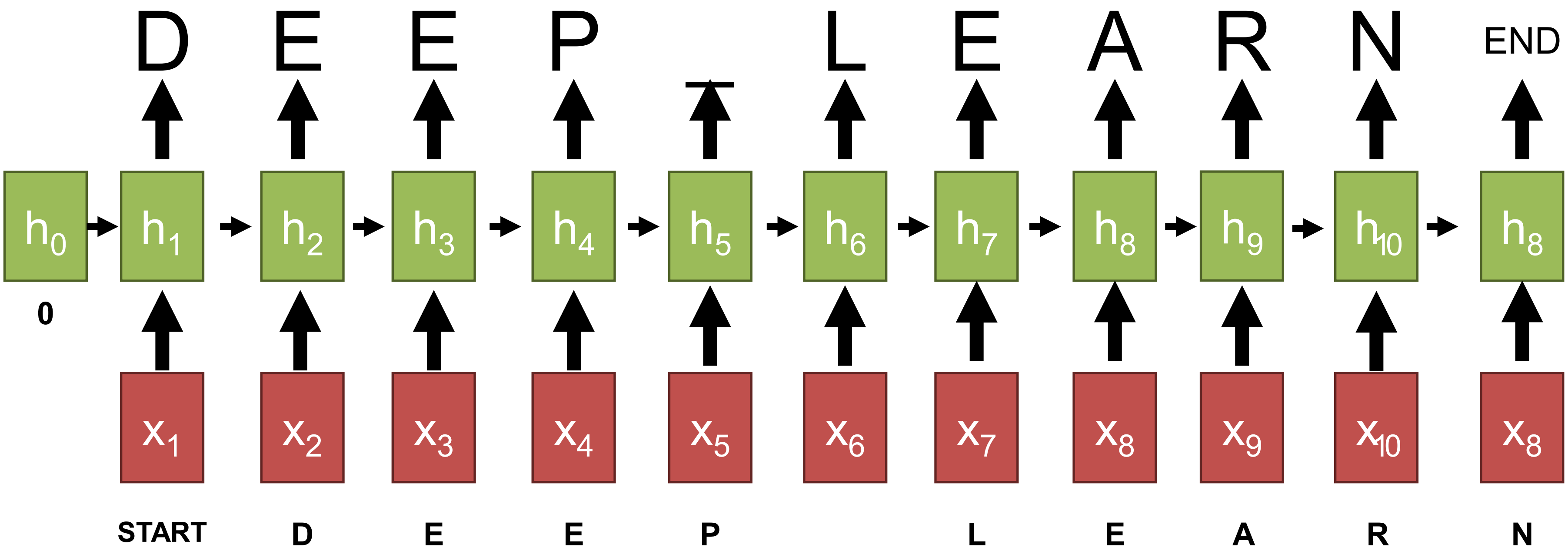
Result credit: A. Karpathy

Nagging Detail #1 – Depth

What happens to really deep networks?

Remember g^n for $g \neq 1$

Gradients explode / vanish



Nagging Detail #1 – Depth

- Typically use more complex methods that better manage gradient flowback (LSTM, GRU)

- General strategy: pass the hidden state to the next timestep as unchanged as possible, only adding updates as necessary

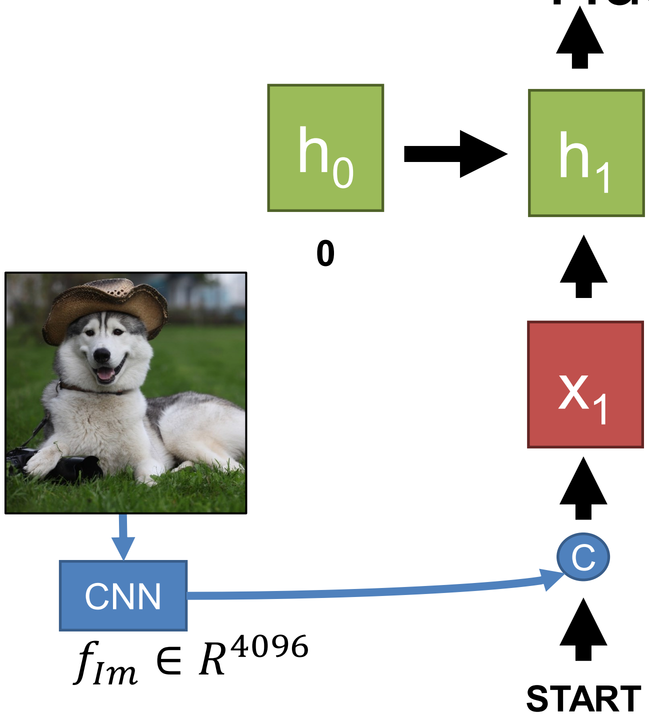
Nagging Detail #2

Lots of captions are in principle possible!

- A dog in a hat
- A dog wearing a hat
- Husky wearing a hat
- Husky holding a camera, sitting in grass
- A dog that’s in a hat, sitting on a lawn with a camera

Nagging Detail #2 – Sampling

Dog (P=0.3), A (P=0.2),
Husky (P=0.15), ….

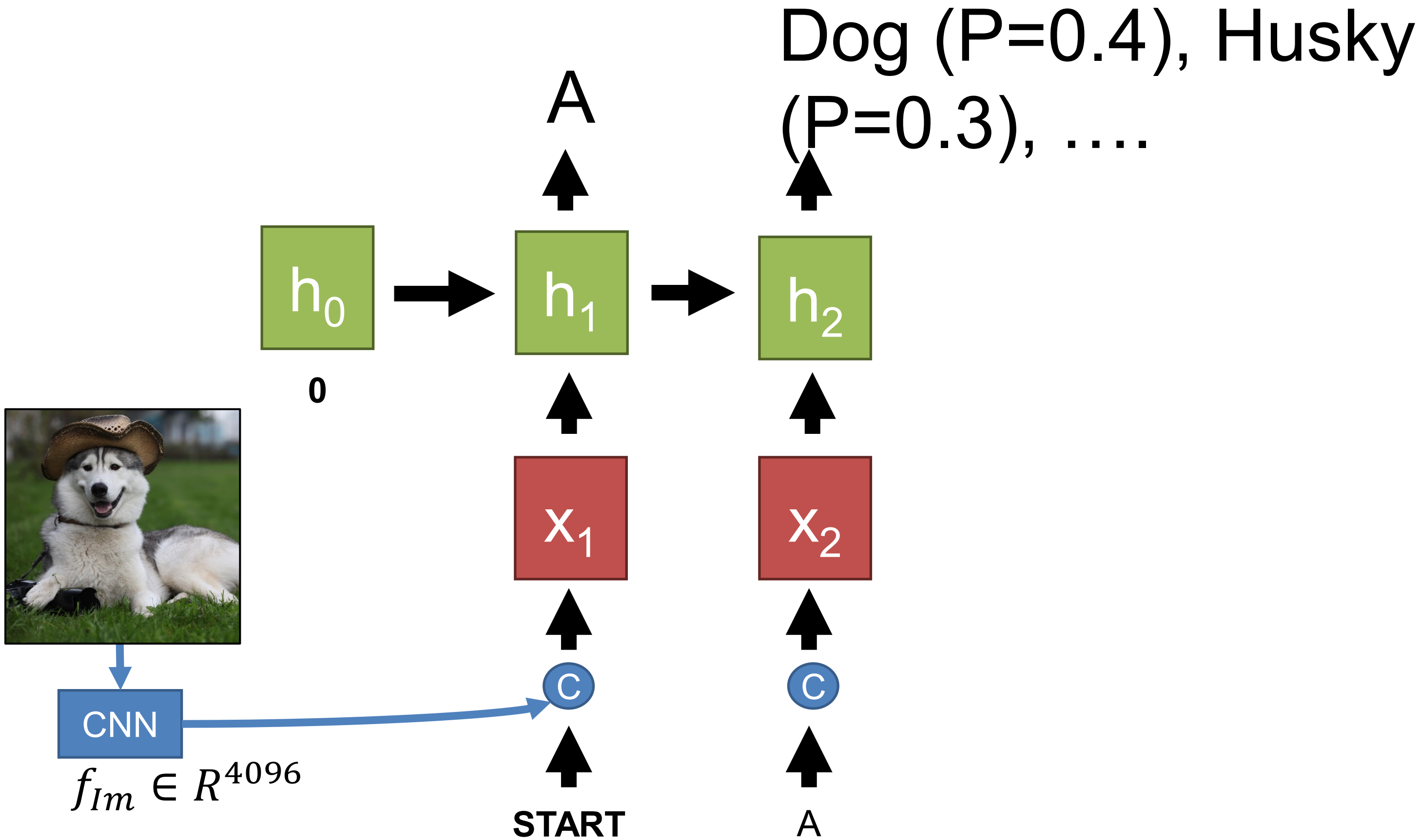


- Pick proportional to probability of each word
- Can adjust "temperature" parameter exp(score/t) to equalize probabilities
- exp(5) / exp(1) → 54.6
- exp(5/5) / exp(1/5) → 2.2

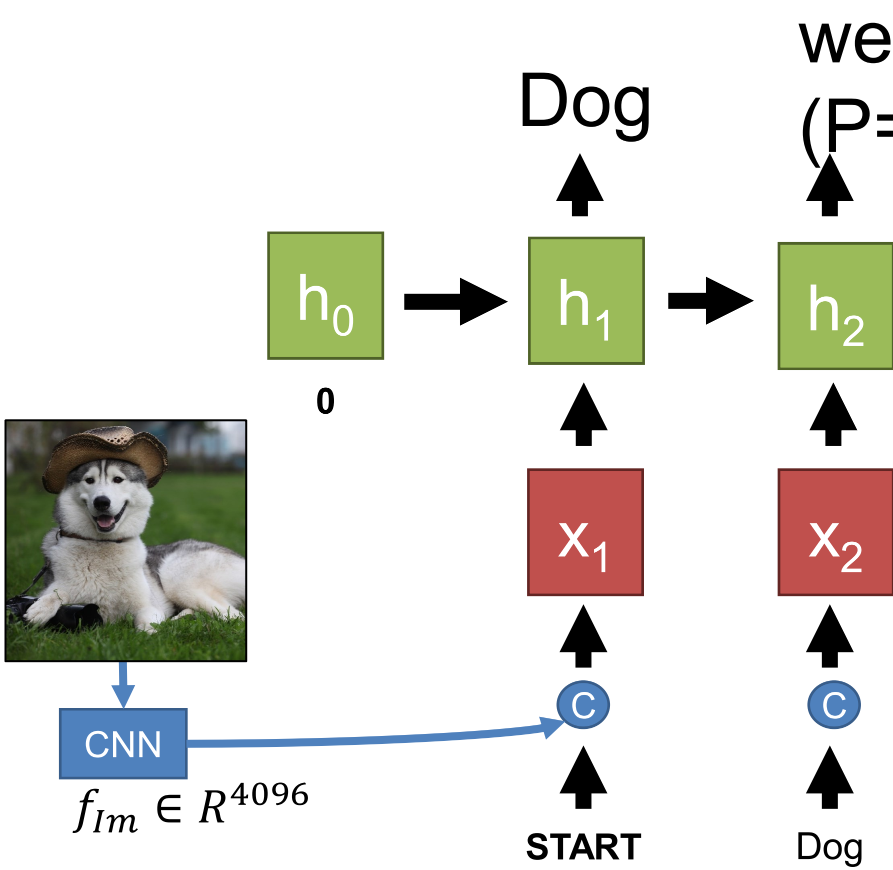
Effect of Temperature

- Train on essays about startups and investing
- Normal Temperature: “The surprised in investors weren’t going to raise money. I’m not the company with the time there are all interesting quickly, don’t have to get off the same programmers. There’s a super-angel round fundraising, why do you can do.”
- Low temperature: “*is that they were all the same thing that was a startup is that they were all the same thing that was a startup is that they were all the same thing that was a startup is that they were all the same*”

Nagging Detail #2 – Sampling



Nagging Detail #2 – Sampling



Each evaluation gives $P(w_i|w_1,...,w_{i-1})$

Can expand a finite tree of possibilities (beam search) and pick most likely sequence

Nagging Detail #3 – Evaluation

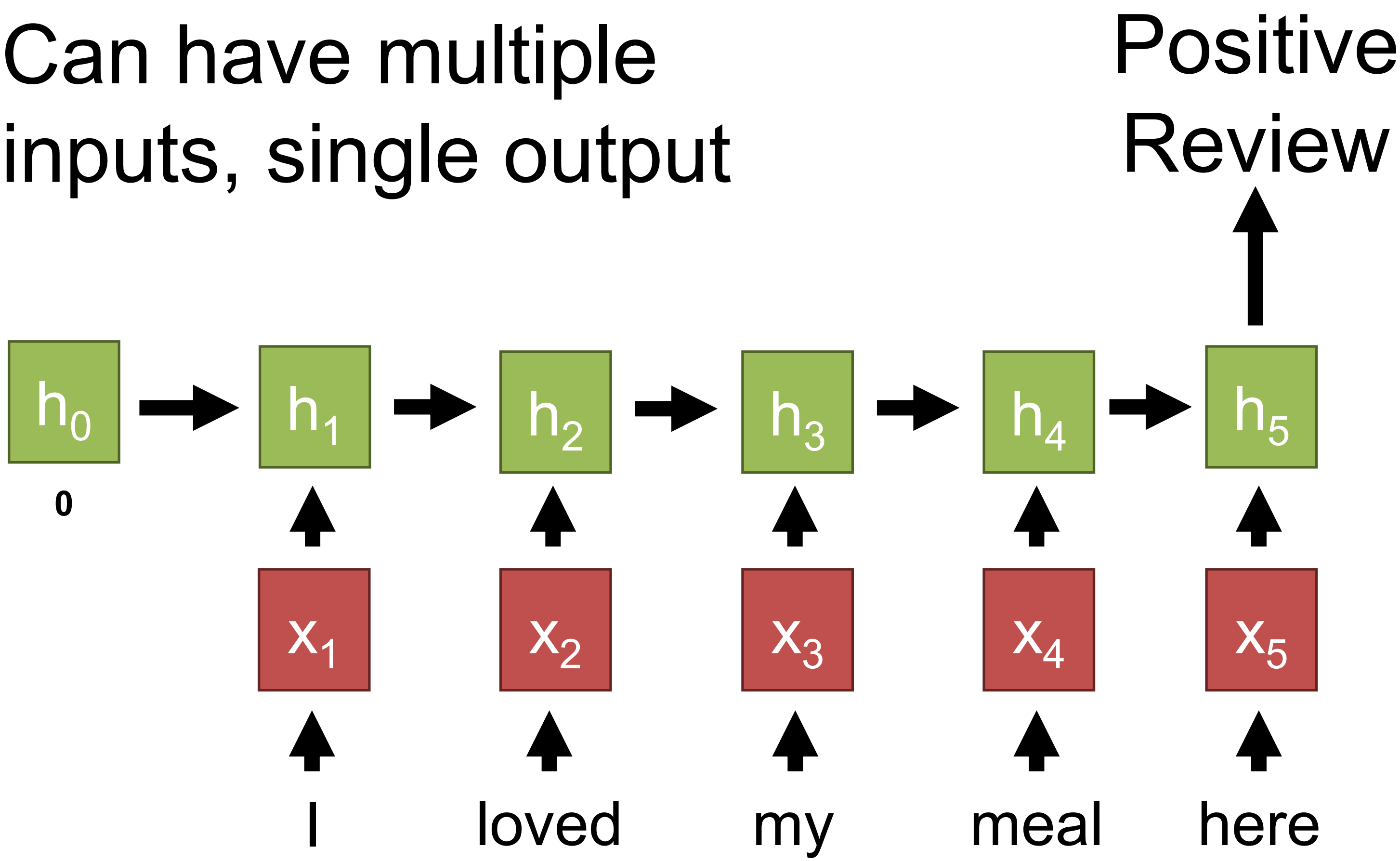
Computer: “A husky in a hat”
Human: “A dog in a hat”

How do you decide?

1) Ask humans. **Why might this be an issue?**
2) In practice: use something like precision (how many generated words appear in ground-truth sentences) or recall. Details very important to prevent gaming (e.g., “A a a a a”)

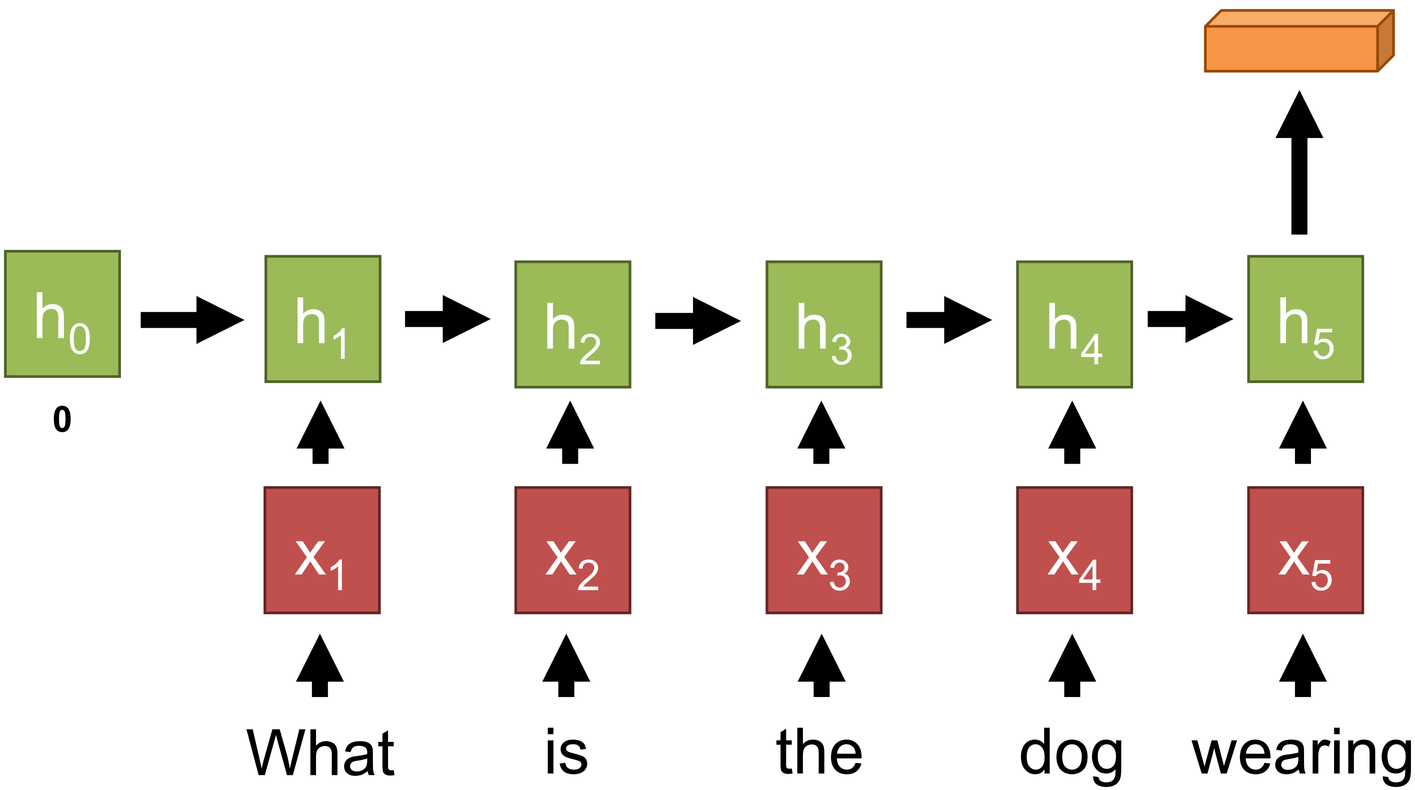
More General Sequence Models

Can have multiple inputs, single output

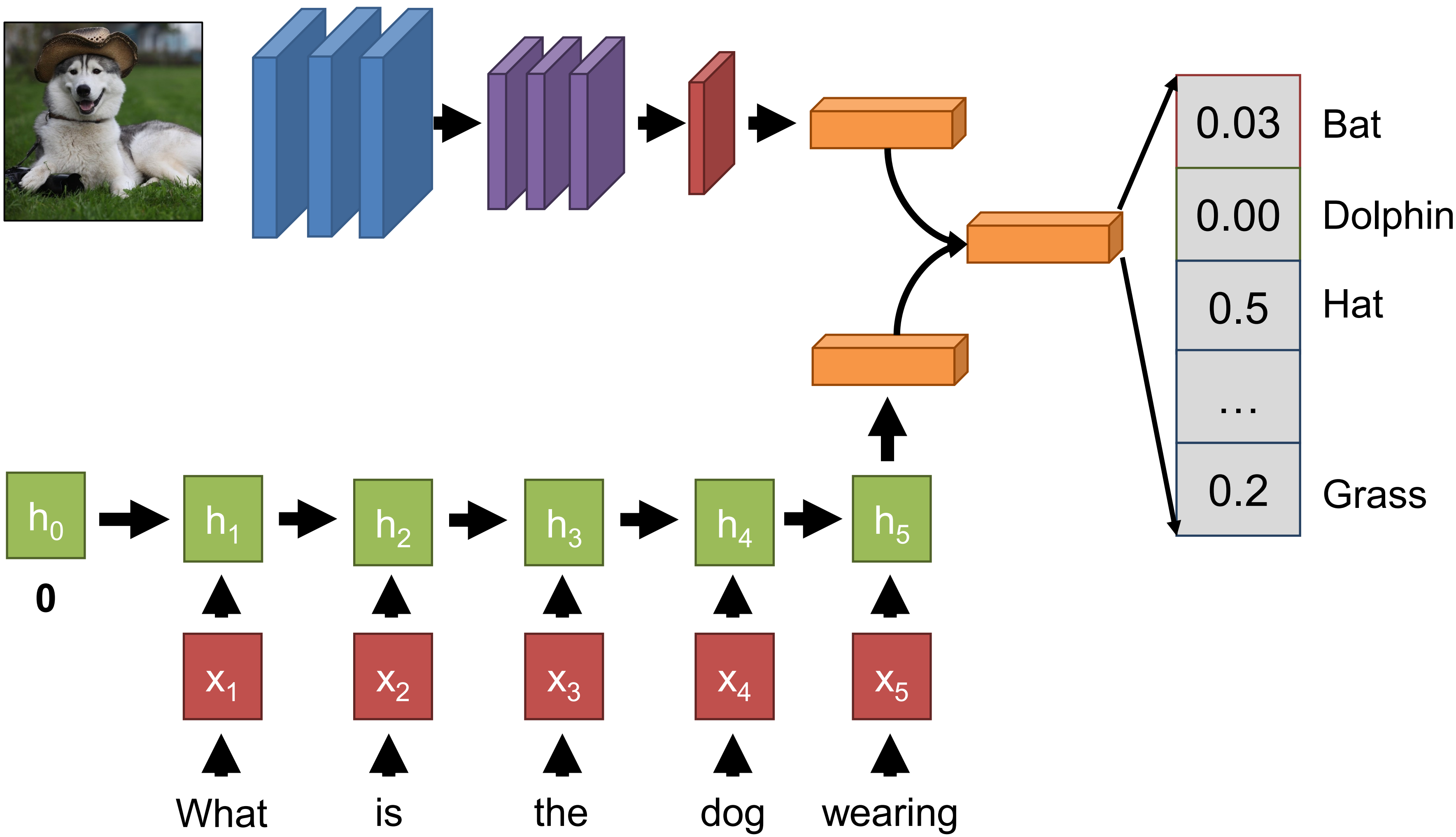


More General Sequence Models

Could be a feature vector!



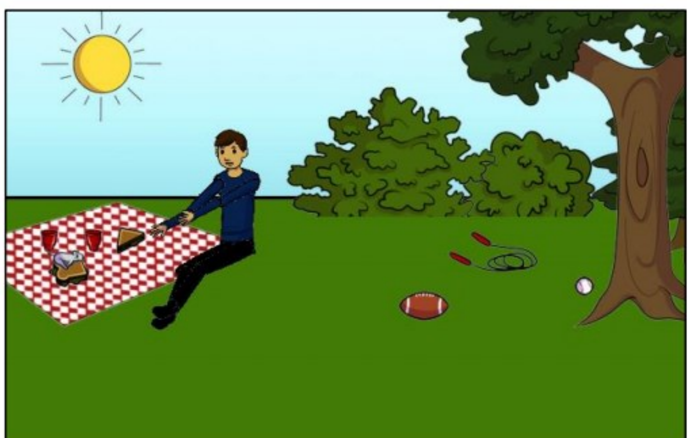
More General Models



Visual Question-Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



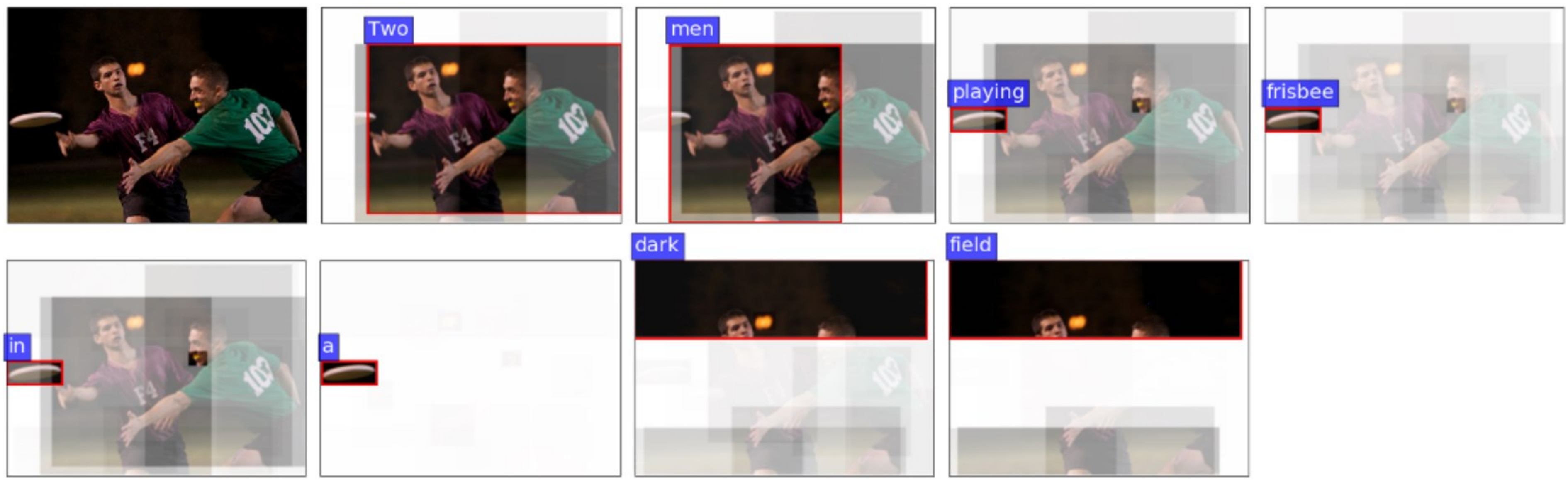
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering. S. Antol, A. Agrawal et al. ICCV 2015

Top-Performing Methods

Top methods now look at objects in the image as opposed to one big image vector.



Two men playing frisbee in a dark field.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al. 2018.

Top-Performing Methods

Question: What color is illuminated on the traffic light? Answer left: green. Answer right: red.

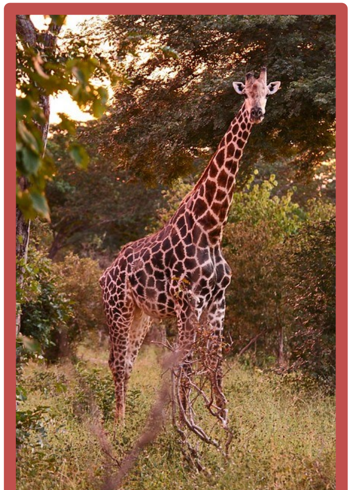
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al. 2018.

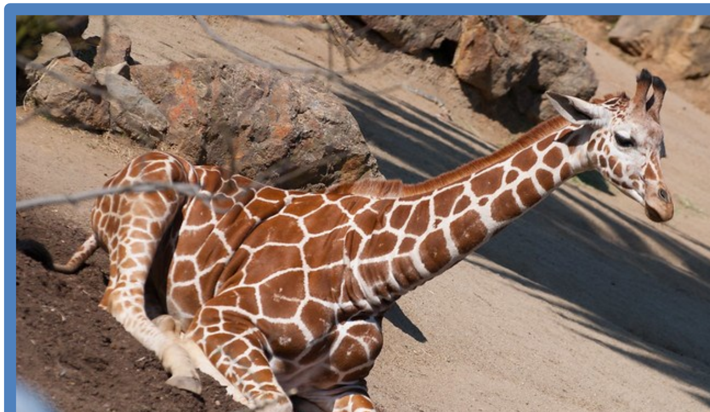
Let's Revisit A Number

- How many 20-word sentences with a vocabulary of 10k words are there really?
- Is it really $(10k)^{20}$? **Why not?**
- Let's look at some giraffes (I swear this is relevant)

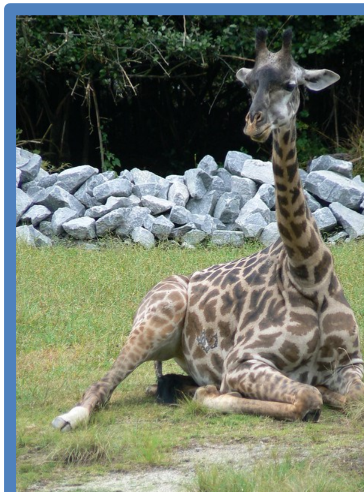
Giraffe example credit: L. Zitnick

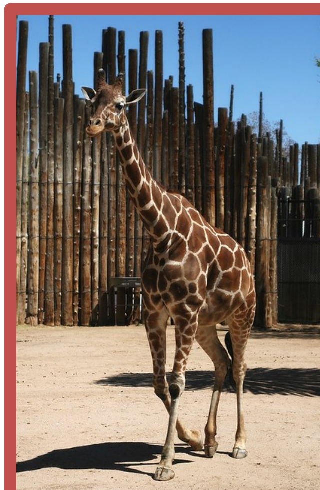
What do Giraffes Do All Day?



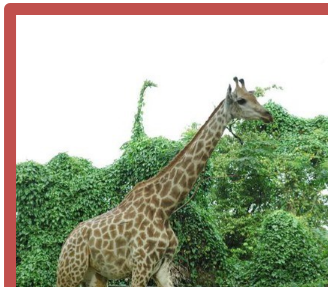


A giraffe sitting and resting





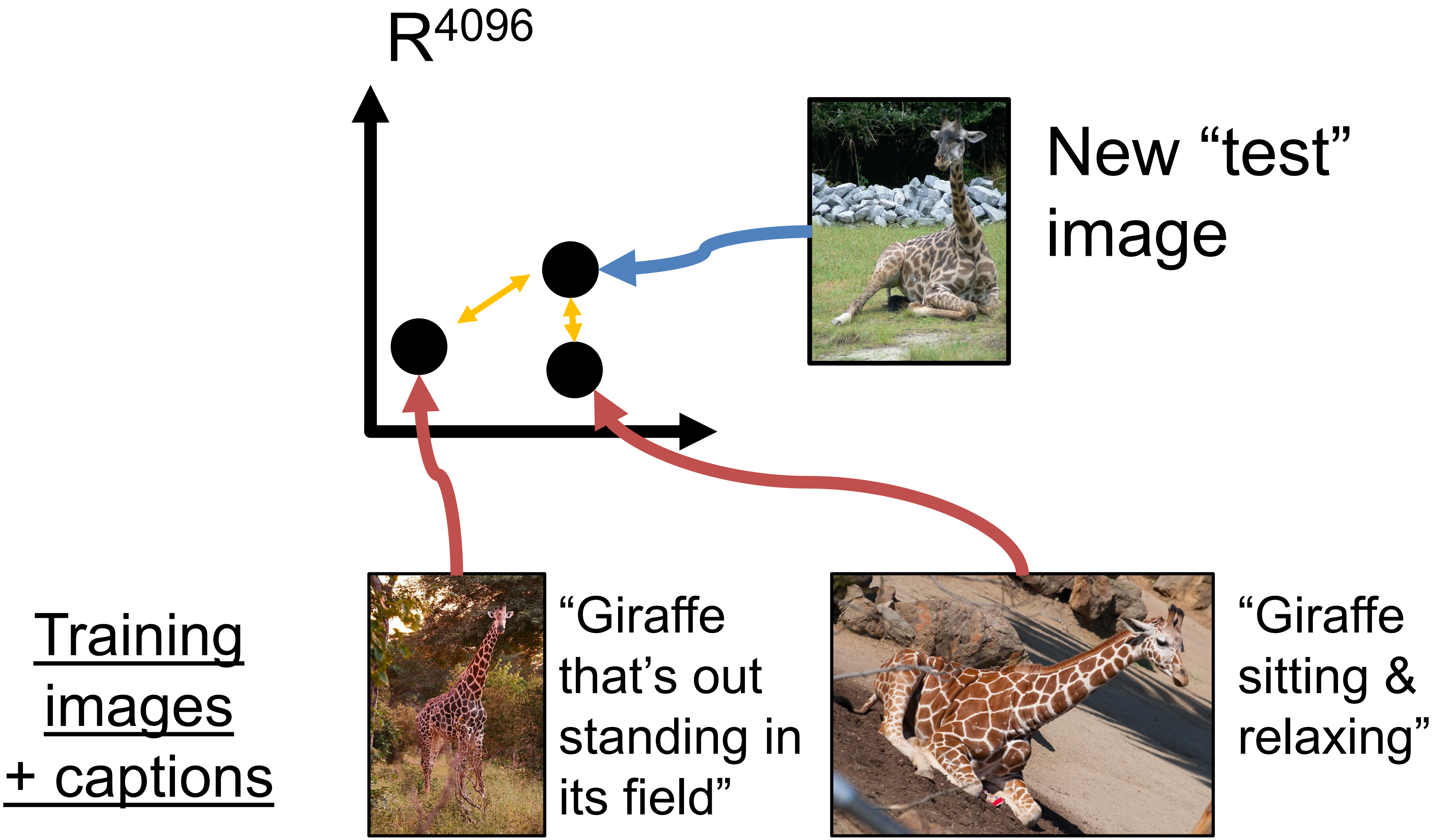
A giraffe grazing in its enclosure



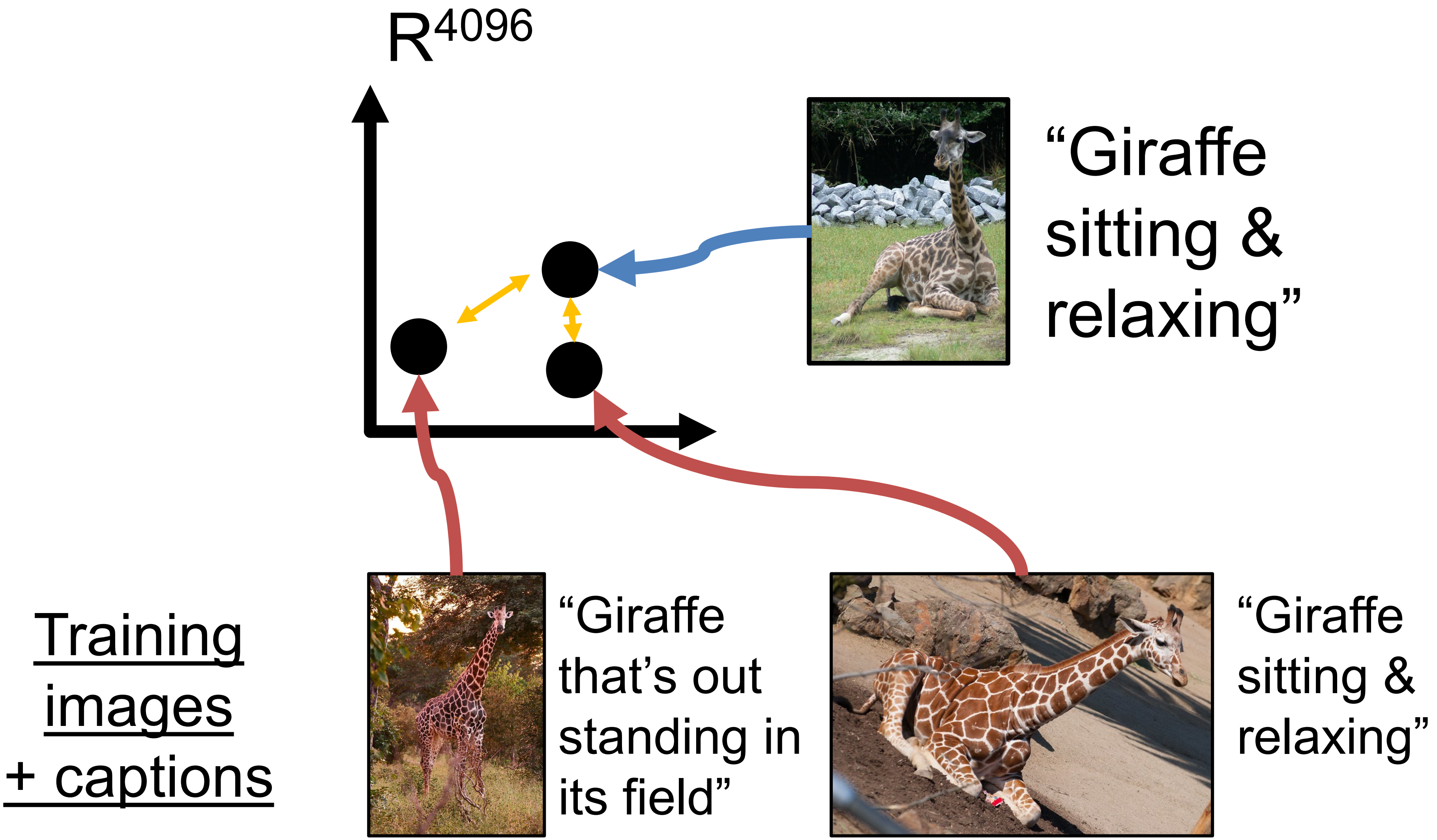
A giraffe wandering around

With apologies to both giraffes and people who study giraffes, I'm sure they're fascinating

Alternate Idea – Retrieval



Alternate Idea – Retrieval



Retrieval Results

A man riding a wave on a surfboard.

A man riding a wave on a surfboard in the ocean.

A person flying a kite in the sky.

A person flying a kite in the sky.

A cat sitting in a bathroom sink.

A black and white cat sitting in a bathroom sink.

Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015

Retrieval Results

A wooden bench in front of a building.

A window display on the front of a building.

A building with a clock on the top.

A clock tower on the top of a building.

The side of a passenger train at a train station.

A bus that is on the side of a road.

Exploring Nearest Neighbor Approaches for Image Captioning. Devlin et al. 2015

Retrieval Results

- In practice: humans don’t like retrieved captions as much
- Can’t generate anything new!
- Works well, but not as well as good systems

Latest in This Space

- *Very* Quick summary
- Great resources include
  - EECS 487 (Lu Wang)
  - EECS 498 (Justin Johnson)
  - Transformers from scratch
    http://peterbloem.nl/blog/transformers

Other Models

Example: English to French

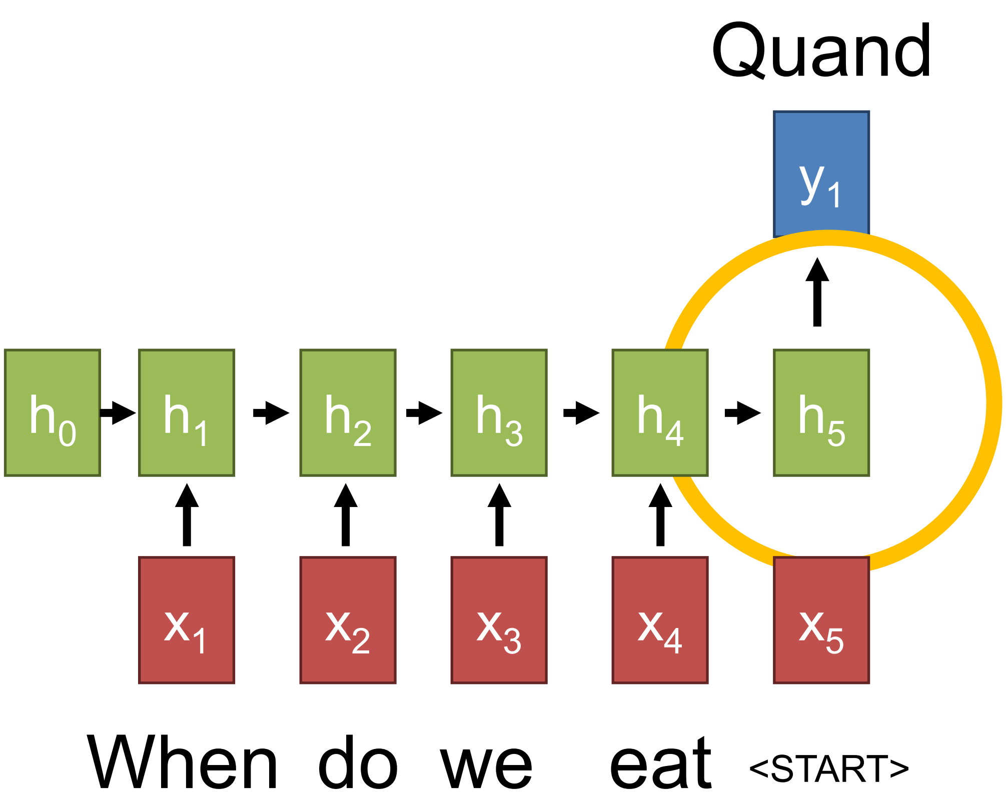
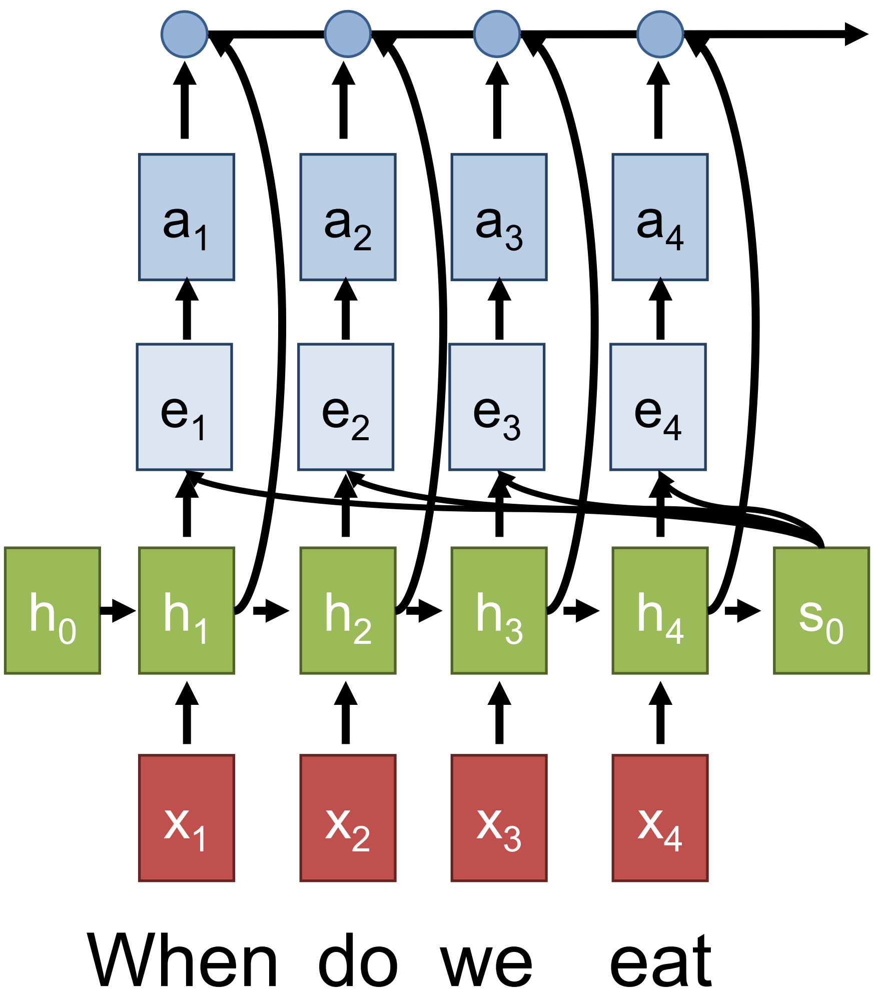


Diagram is remake of one from J. Johnson

Suppose you're in charge of the hidden state. What's the problem?

Have to keep track of everything! Very tedious if the sequence is long.

Self-Attention



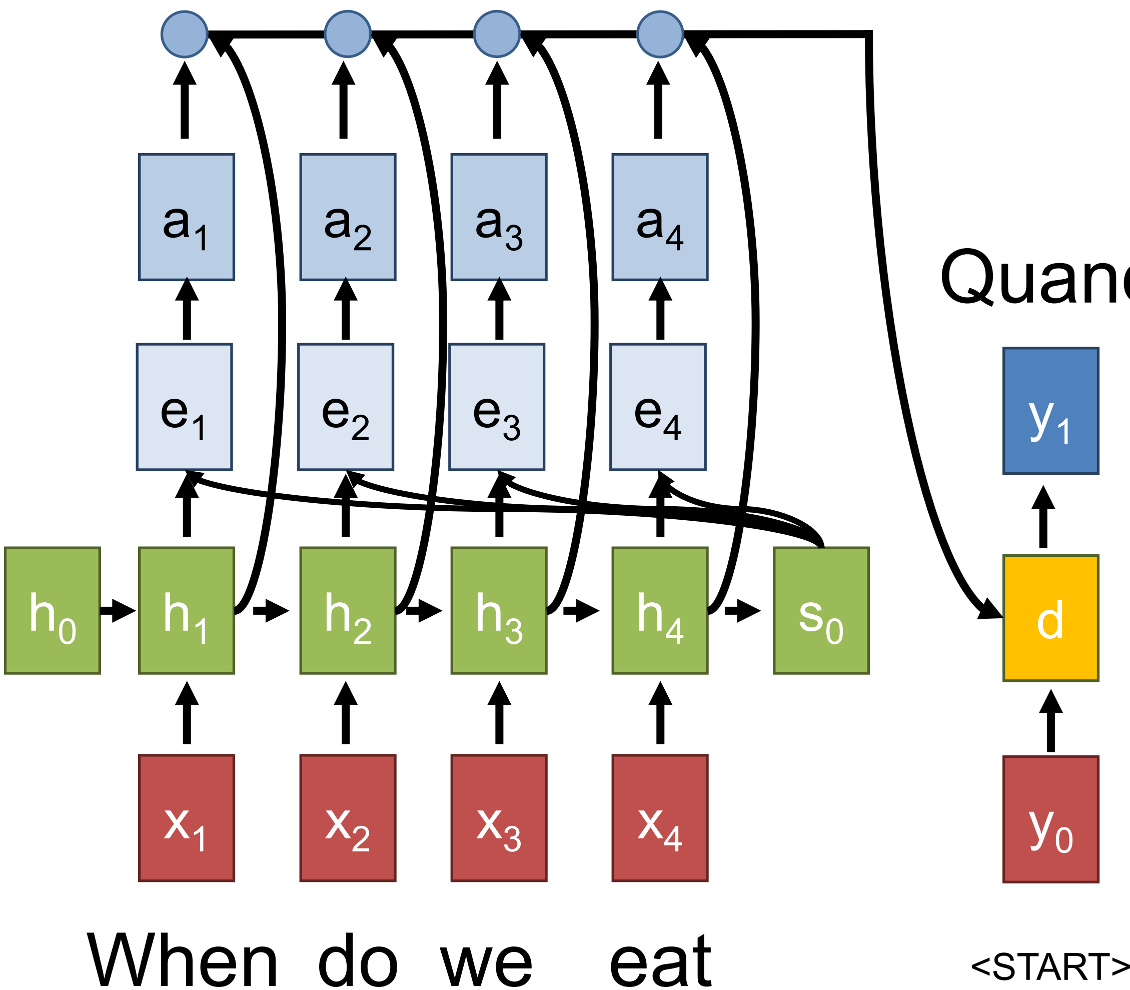
Output: $\sum_i a_i \boldsymbol{h_i}$

Attention: softmax over e_i

Alignment Score: $e_i = \boldsymbol{h}_i^T \boldsymbol{s}_0$

Diagram is remake of one from J. Johnson

Self-Attention



Output given to decoder that handles making predictions.

Diagram is remake of one from J. Johnson

Transformers

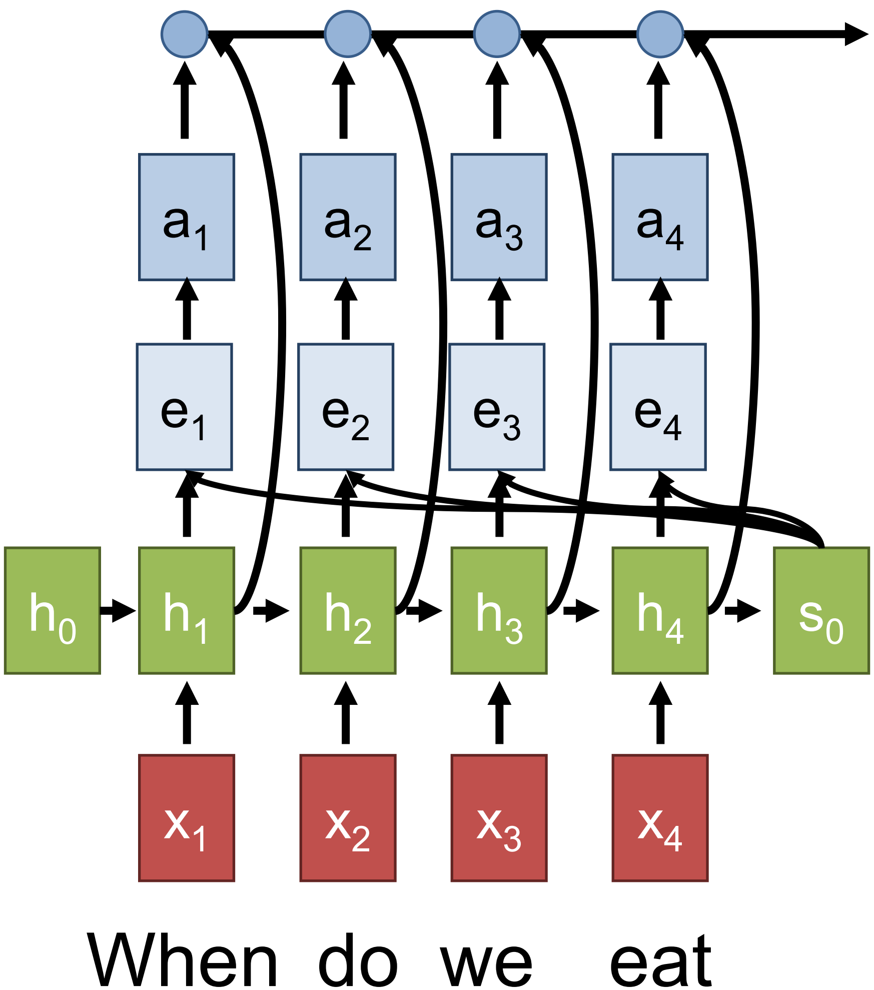


Diagram is remake of one from J. Johnson

Old Output: $\sum_i a_i \boldsymbol{h_i}$

New Output: $\sum_i a_i \left(\boldsymbol{V h_i}\right)$

Vector doesn't do it all

Attention: softmax over e_i

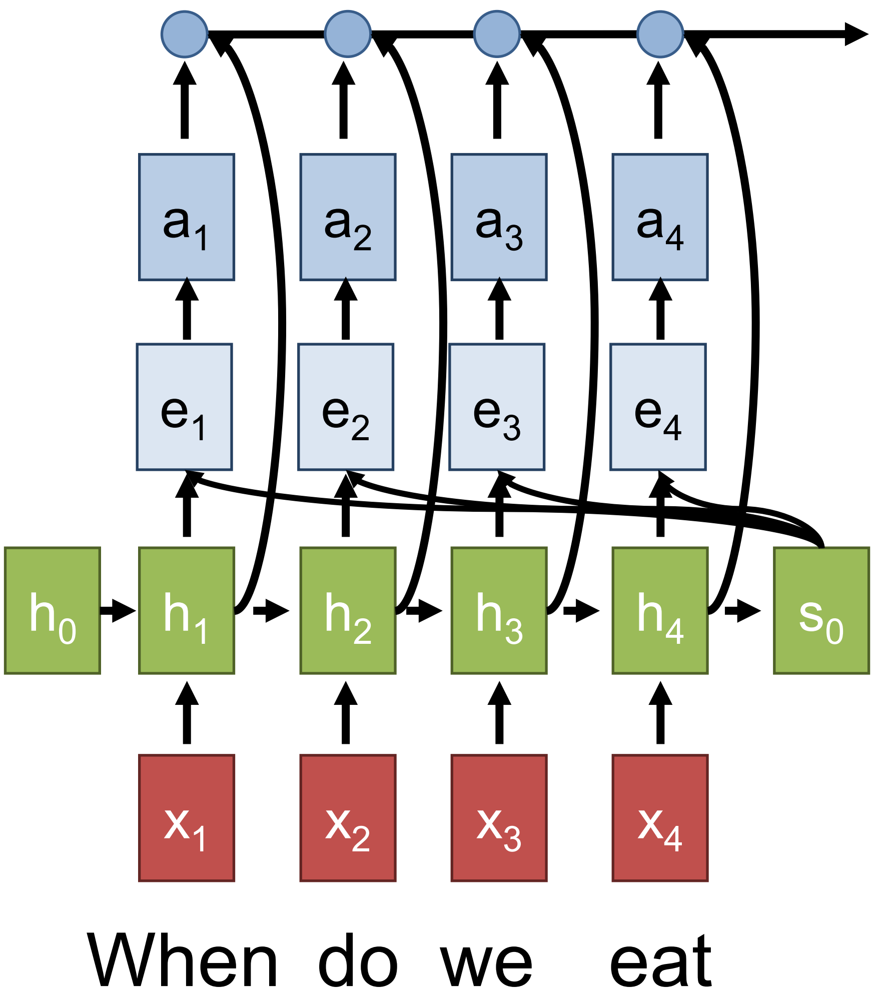
Old Alignment: $e_i = \boldsymbol{h}_i^T \boldsymbol{s}_0$

New: calculate e_i with projection of vectors $\boldsymbol{h}_i^T \boldsymbol{K} \boldsymbol{s}_0$

Vector doesn't do it all

New: scale dot product between F-D vectors by $\sqrt{F}$

Transformers



Does the alignment score know about position?

Alignment Score: $e_i = \boldsymbol{h}_i^T \boldsymbol{s}_0$

Solution: add vectors that are function of location

Diagram is remake of one from J. Johnson

Positional Encodings

Set of sinusoids of different frequencies. Really effective trick for encoding locations for networks

$$f(p, 2i) = \sin\left(\frac{p}{10000^{2i/d}}\right) \quad f(p, 2i+1) = \cos\left(\frac{p}{10000^{2i/d}}\right)$$

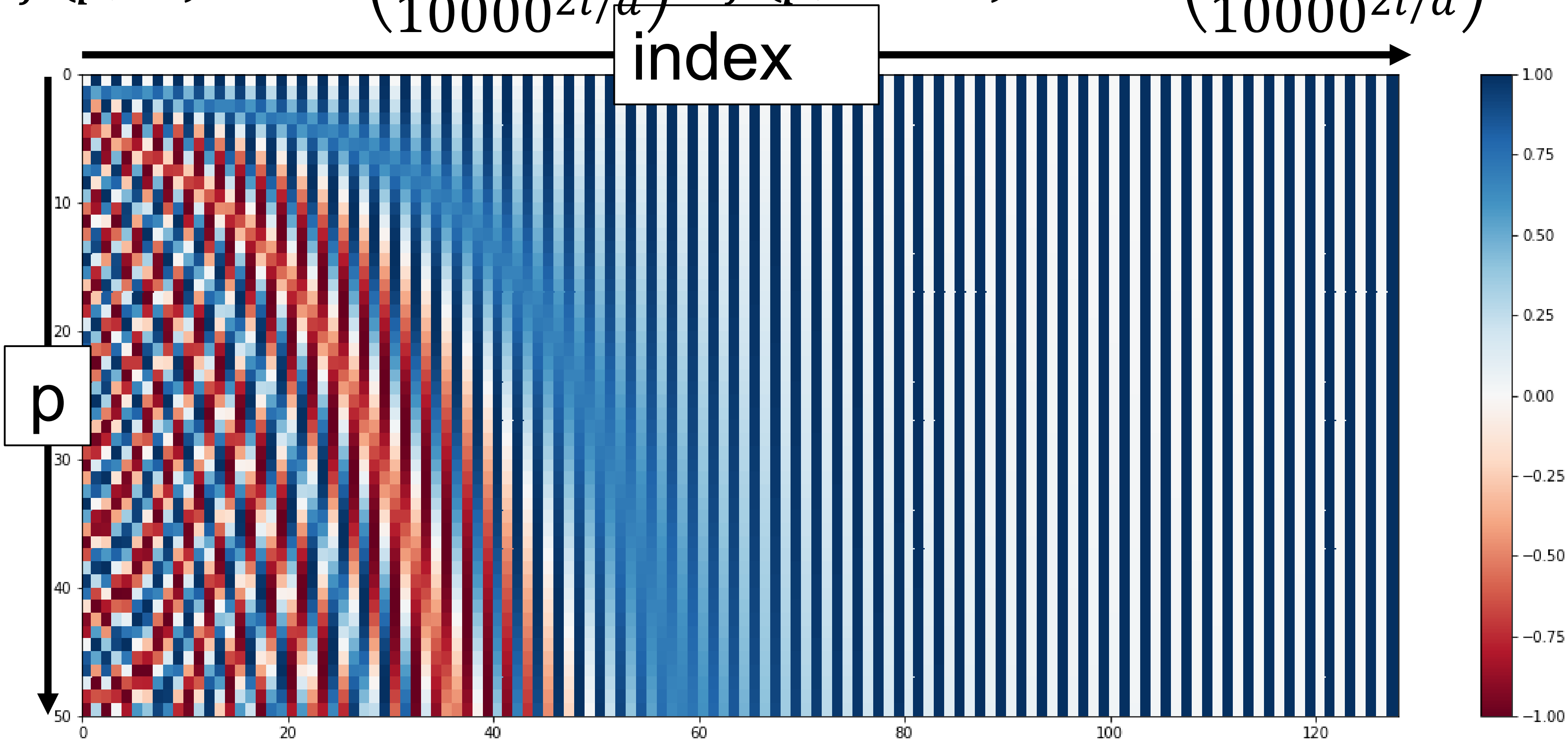


Diagram: Amirhoseein Kazemjejad,
https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

Large-scale Language Models

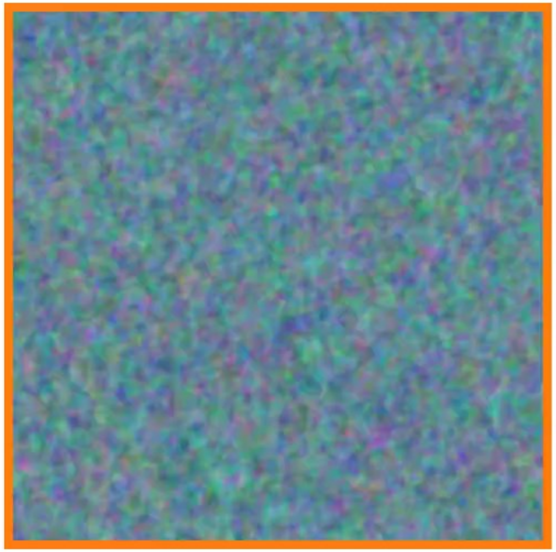
- Can train these models at huge scale
- GPT-3:
  - 410 billion tokens of training data
  - 17 billion parameters
- Worth reading: “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜” Bender et al.

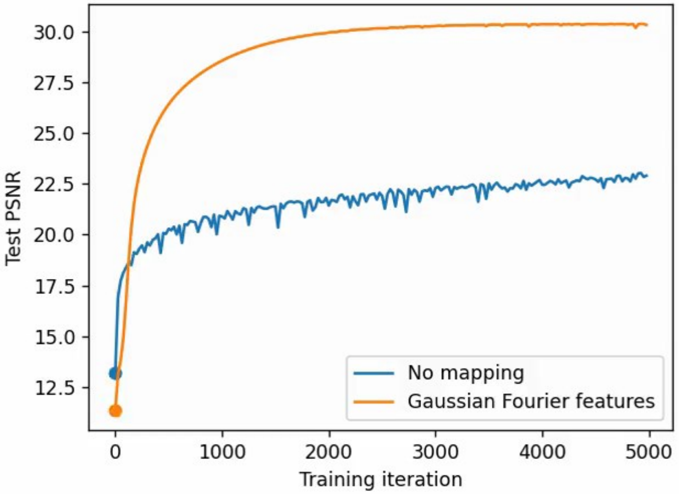
Positional Encodings Elsewhere

Learn network mapping from (x,y) -> (r,g,b)

Learned on (x,y)

w/positional encodings)





Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains, Tancik et al. NeurIPS 2020. See also Implicit Neural Representations with Periodic Activation Functions, Sitzmann et al. NeurIPS 2020.

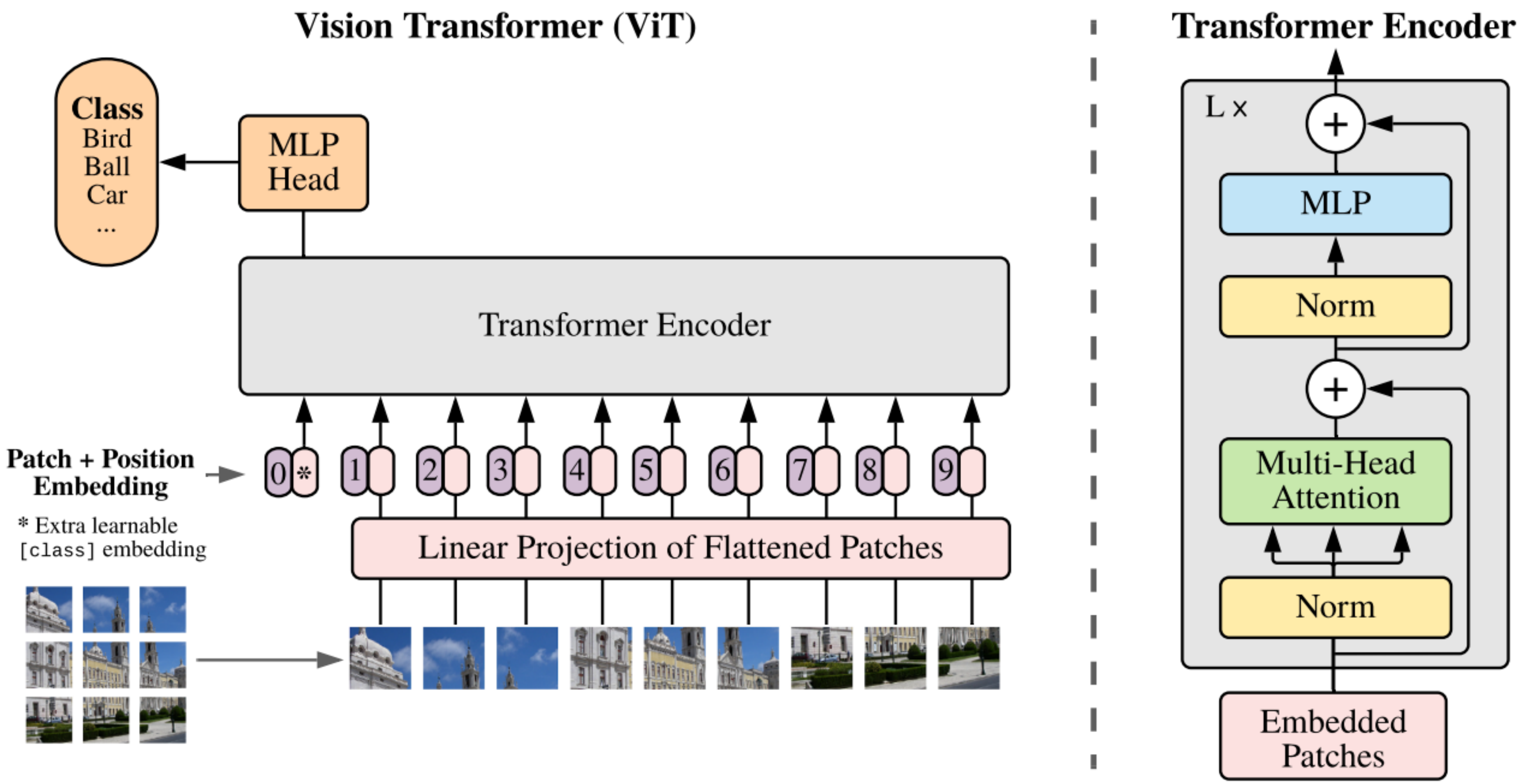
How Might We Use a Transformer?

Let’s plug an image into a transformer

Break into patches, treat like words

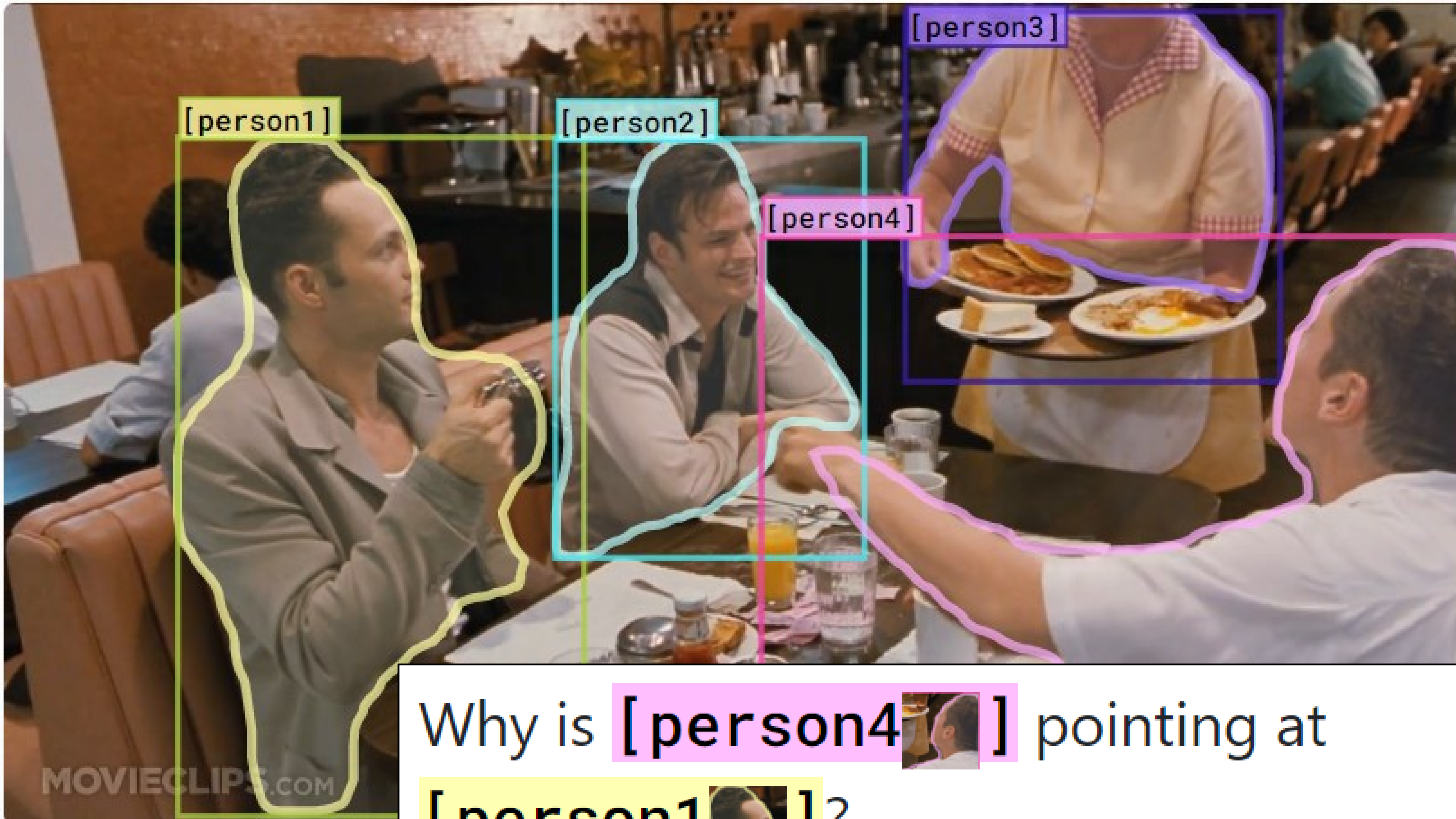
Vision Transformer

Key idea: put in sequence of image tokens



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al. ICLR 2021.

VCR

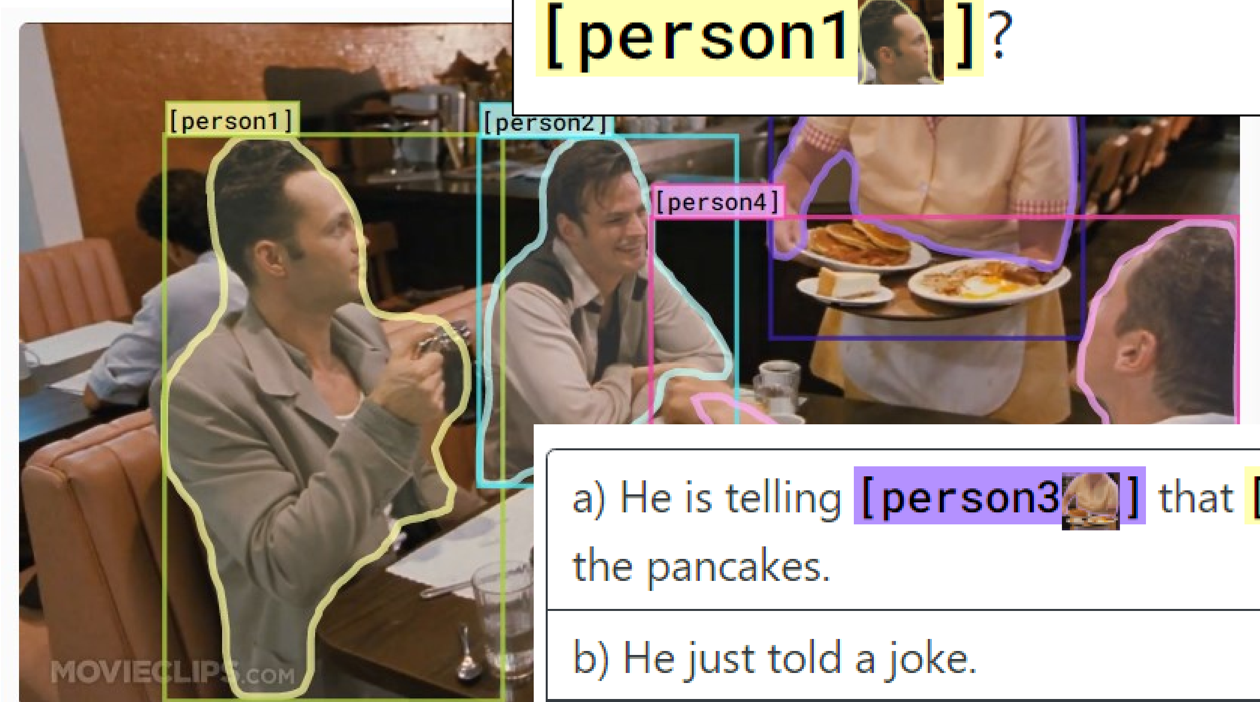


Why is [person4] pointing at [person1]?

From Recognition to Cognition: Visual Commonsense Reasoning. Zellers et al. CVPR 2019

VCR

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.

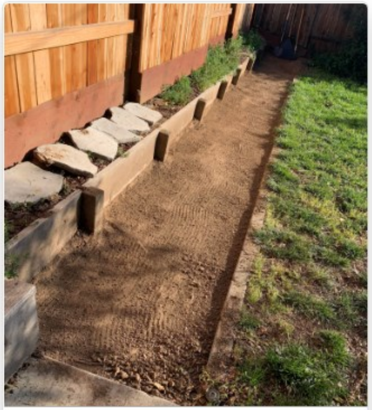
From Recognition to Cognition: Visual Commonsense Reasoning. Zellers et al. CVPR 2019

RedCaps


r/breadit: chocolate babka!


r/woodworking: rabbet cutting - what would you do?


r/itookapicture: itap of a soon to be sunset in cozumel, mexico.


r/gardening: i want to build a large planter box. any suggestions?


r/guineapigs: pumpkin peeking out

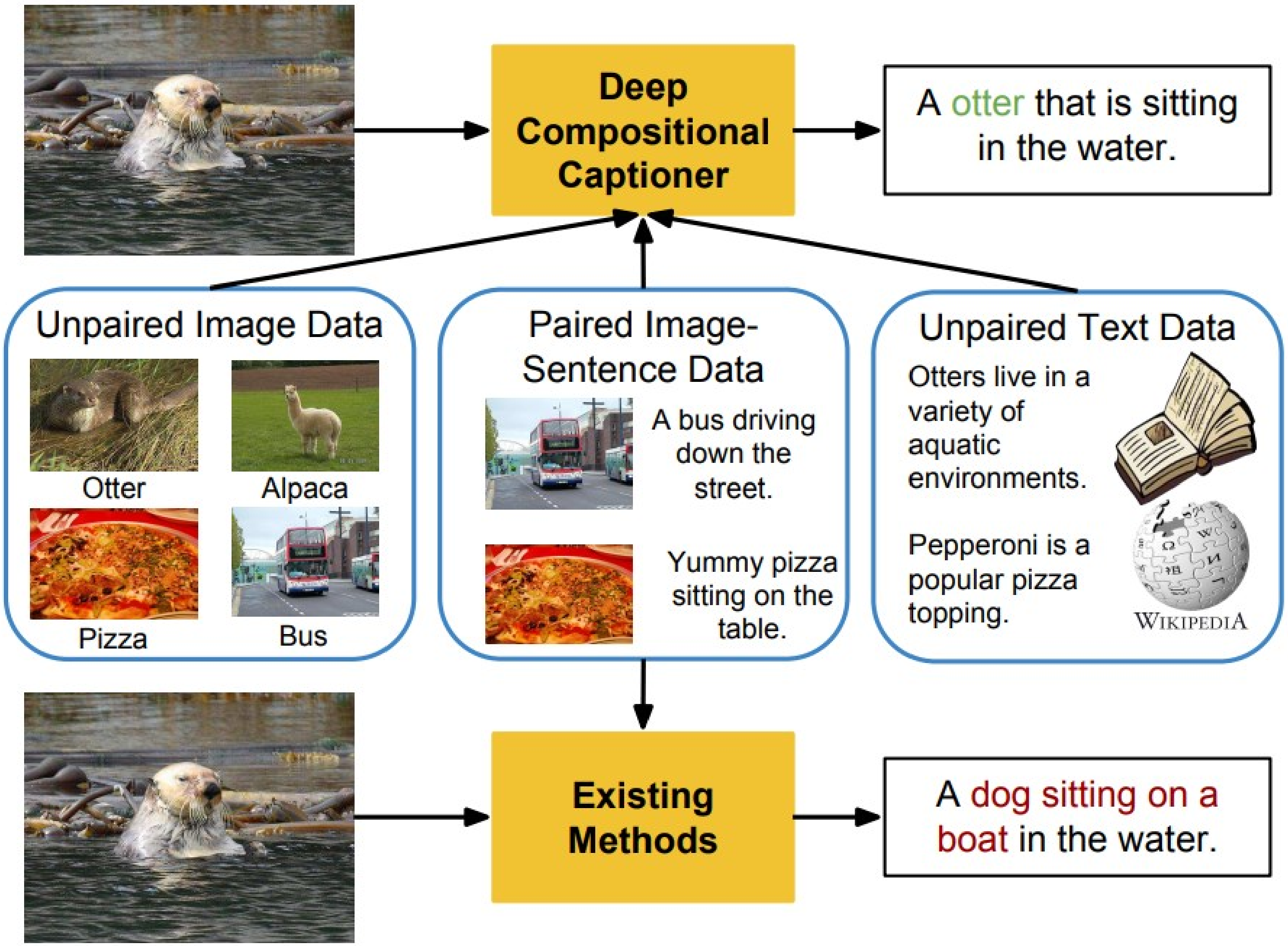
redcaps.xyz

RedCaps: web-curated image-text data created by the people, for the people. Desai et al. NeurIPS 2021.

Some Concluding Thoughts

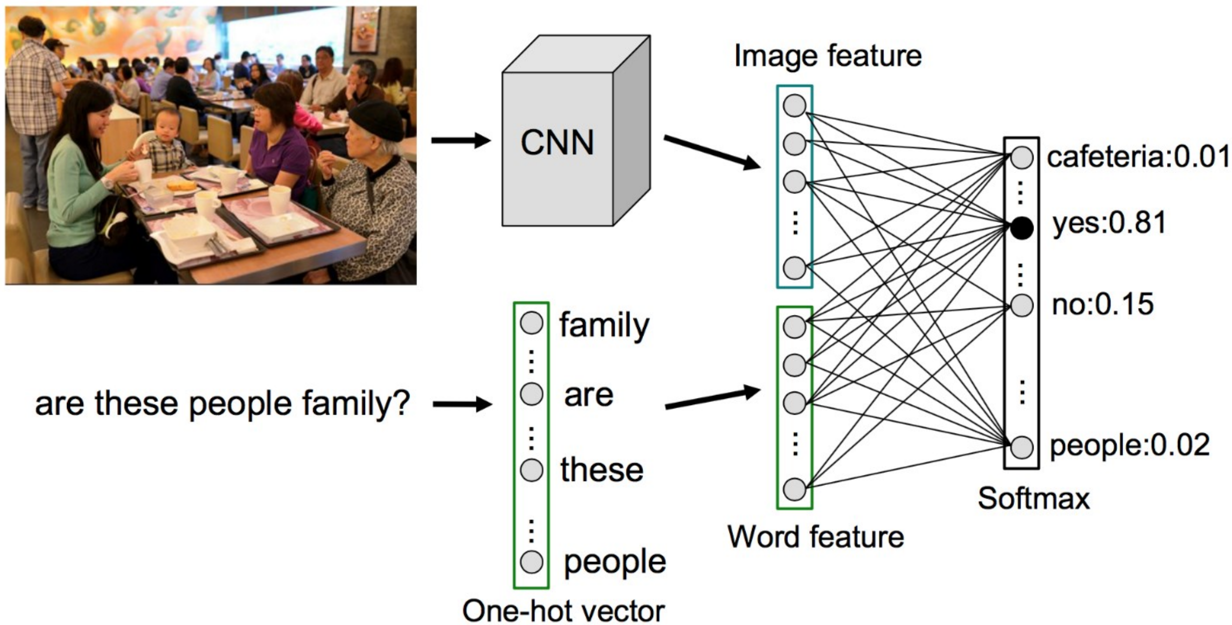
- Getting this right is really hard!
- Deep learning is trying to do solve any problem you pose with as little effort as possible.
- A lot of this has to do with data and people

Novel Captions



Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data.
L. Hendricks et al. CVPR 2016

Simple Baseline for VQA



- Construct a vocabulary of 5000 most frequent answers
- Extract all the information from the image, I
  - Construct an image representation using a CNN
- Represent the question, Q with BoW
- Compute distribution of answers, $P(A|Q, I)$

[Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).]

Slide credit: T. Gupta

Qualitative Results

Question: what are they doing
Predictions:

playing baseball (score: 10.67 = 2.01 [image] + 8.66 [word])
baseball (score: 9.65 = 4.84 [image] + 4.82 [word])
grazing (score: 9.34 = 0.53 [image] + 8.81 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)
Based on word only: playing wii (10.62), eating (9.97), playing frisbee (9.24)

Question: how many people inside
Predictions:

3 (score: 13.39 = 2.75 [image] + 10.65 [word])
2 (score: 12.76 = 2.49 [image] + 10.27 [word])
5 (score: 12.72 = 1.83 [image] + 10.89 [word])

Based on image only: umpire (4.85), baseball (4.84), batter (4.46)
Based on word only: 8 (11.24), 7 (10.95), 5 (10.89)

Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

Slide credit: T. Gupta

Qualitative Results

Question: which brand is the laptop
Predictions:
apple (score: 10.87 = 1.10 [image] + 9.77 [word])
dell (score: 9.83 = 0.71 [image] + 9.12 [word])
toshiba (score: 9.76 = 1.18 [image] + 8.58 [word])

Based on image only: books (3.15), yes (3.14), no (2.95)
Based on word only: apple (9.77), hp (9.18), dell (9.12)

- Language prior prunes the answer space significantly

Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

Slide credit: T. Gupta

Quantitative Evaluation

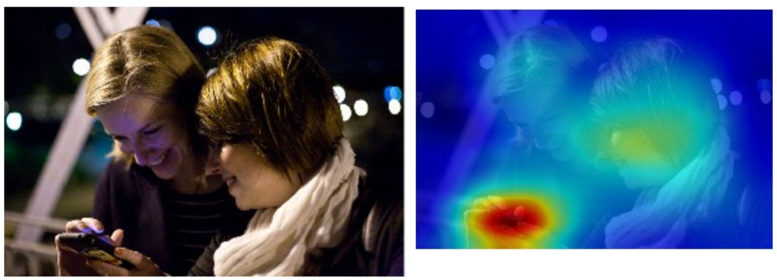
	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
IMG [2]	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BOW [2]	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
BOWIMG [2]	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTMIMG [2]	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CompMem [6]	52.62	78.33	35.93	34.46	-	-	-	-
NMN+LSTM [1]	54.80	77.70	37.20	39.30	-	-	-	-
WR Sel. [13]	-	-	-	-	60.96	-	-	-
ACK [16]	55.72	79.23	36.13	40.08	-	-	-	-
DPPnet [11]	**57.22**	80.71	37.24	41.69	**62.48**	80.79	38.94	52.16
iBOWIMG	55.72	76.55	35.03	42.62	61.68	76.68	37.05	54.44

Evaluated on the VQA dataset – although now results are quite a bit higher

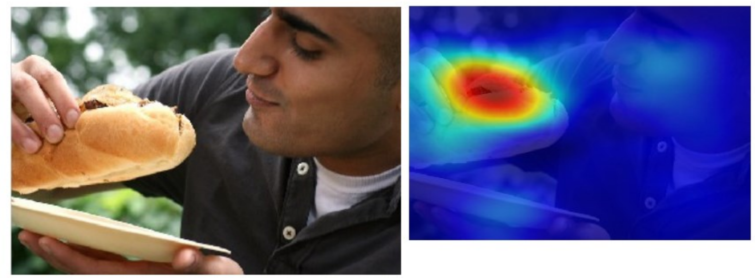
Slide credit: T. Gupta

Does the model learn to localize?

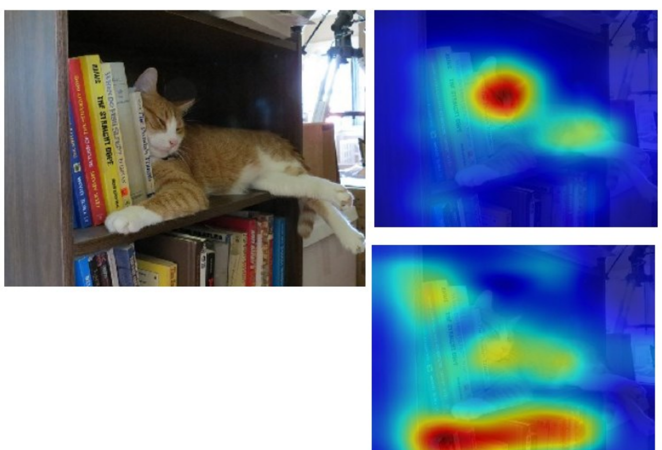
Class Activation Mapping applied to VQA Baseline



Question: What are they doing?
Prediction: texting (score: 12.02=3.78 [image] + 8.24 [word])
Word importance: doing(7.01) are(1.05) they(0.49) what(-0.3)



Question: What is he eating?
Prediction: hot dog (score: 13.01=5.02 [image] + 7.99 [word])
Word importance: eating(4.12) what(2.81) is(0.74) he(0.30)



Question: Is there a cat?
Prediction: yes (score: 11.48 = 4.35 [image] + 7.13 [word])
word importance: is(2.65) there(2.46) a(1.70) cat(0.30)

Question: Where is the cat?
Prediction: shelf (score: 10.81 = 3.23 [image] + 7.58 [word])
word importance: where(3.89) cat(1.88) the(1.79) is(0.01)

Slide credit: T. Gupta

Recent Developments

Can balance data to make things difficult

Who is wearing glasses?

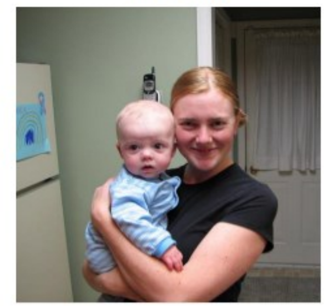
man

woman

Where is the child sitting?

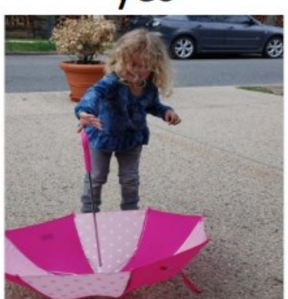
fridge

arms

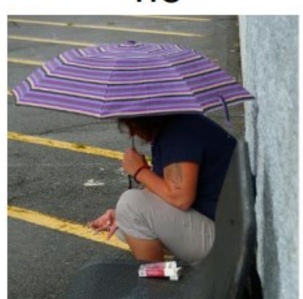


Is the umbrella upside down?

yes



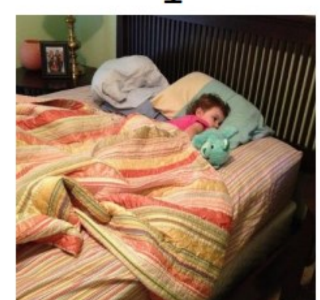
no



How many children are in the bed?

2

1



Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. Goyal et al. 2017