Role of Language in Vision

Computer Vision

CS 543 / ECE 549

University of Illinois

Tanmay Gupta

Today's class: Role of Language in Vision

- Part I:
  - Moving from Classification to Embedding Models for recognition
- Part II:
  - Language representation
- Part III: Hot topics in Vision-Language Research
  - Phrase Localization
  - Visual Question Answering
  - Image Captioning

PART I

MOVING FROM CLASSIFICATION TO EMBEDDING MODELS

Recognition as Image Classification

- Many visual recognition tasks posed as k-way classification with **exclusive** categorical labels

Dog Breed Classification

Classifier →

0.1	Pug
0.2	Bulldog
0.6	Husky
0.05	Poodle
0.05	Dalmatian

Recognition as Image Classification

- Many visual recognition tasks posed as k-way classification with **exclusive** categorical labels

Modified Dog Breed Classification

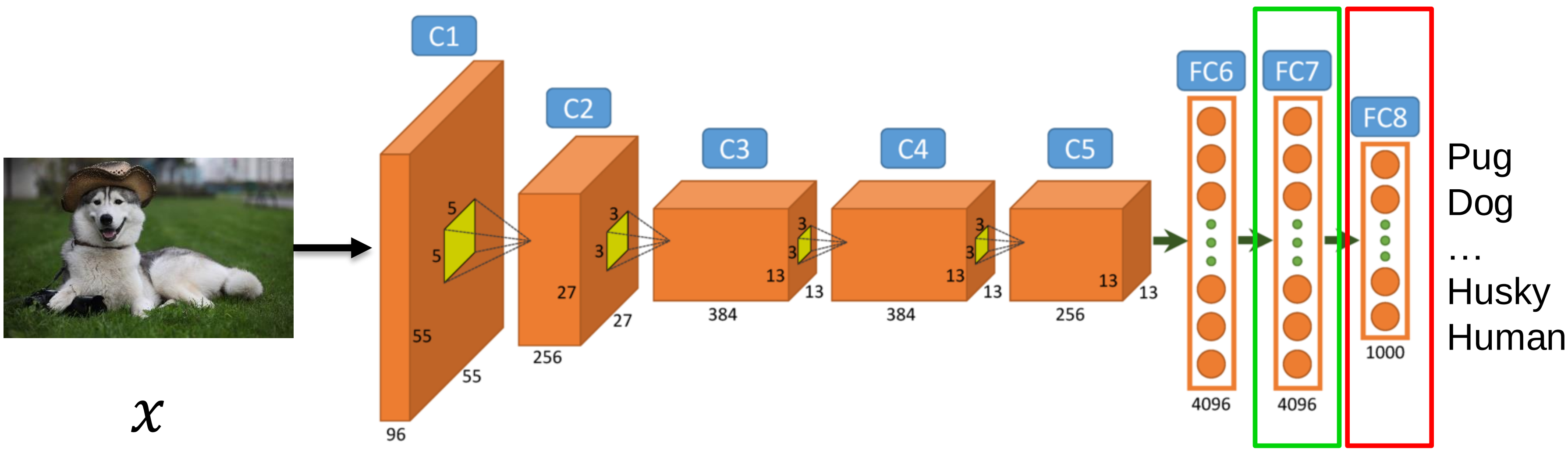
Classifier

0.1	Pug
0.2	Dog
0.6	Husky
0.05	Human
0.05	Animal

Recognition as Image Classification

- Many visual recognition tasks posed as k-way classification with **exclusive** categorical labels

Modified Dog Breed Classification



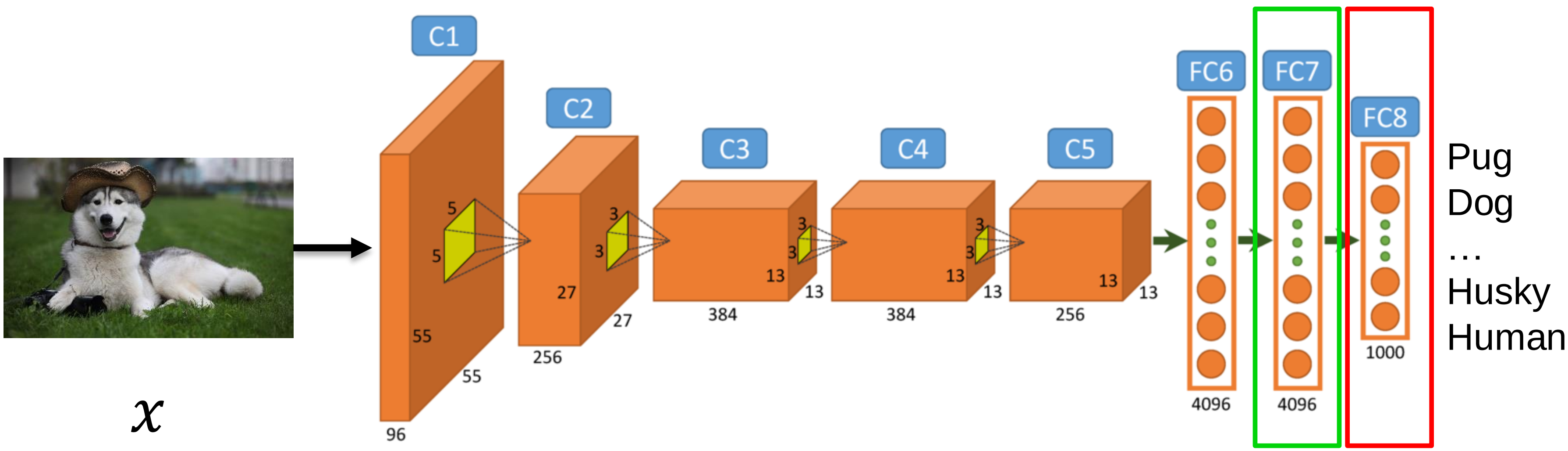
Output of FC7 Layer = Image Representation, $\phi(x)$

FC8 Weights, W act as linear classifiers $s(x, y) = w_y \cdot \phi(x)$

Recognition as Image Classification

- Many visual recognition tasks posed as k-way classification with **exclusive** categorical labels

Modified Dog Breed Classification



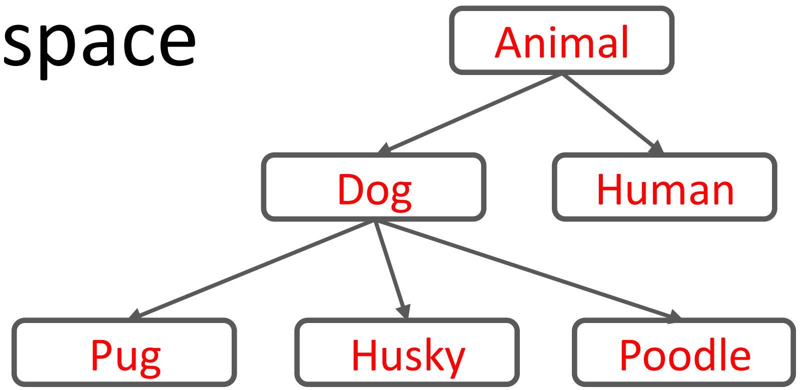
$$s(x, y) = w_y \phi(x) \qquad P(y_i|x) = \frac{e^{s(x,y_i)}}{\sum_y e^{s(x,y)}}$$

Limitations of the classification approach

- Hard to scale to large number of classes

$$P(y_i|x) = \frac{e^{s(x,y_i)}}{\sum_y e^{s(x,y)}}$$ O(#classes)

- Ignores structure in label space
  - Hypernyms (Dog-Husky)
  - Co-Hyponyms (Pug-Husky)



- Ignores additional information about classes available in the form of text

Scalability **Structure in label space** **External Knowledge**

$$\underbrace{s(x, y) = w_y \phi(x)}_{\text{}} \qquad P(y_i|x) = \frac{e^{s(x,y_i)}}{\sum_y e^{s(x,y)}}$$

Compatibility / Scoring Function

- Score is enough to make a prediction
  - Eliminate probability computation during training
  - Consider only relative ranking of subset of classes
- Design compatibility functions that encode structure in the label space

Compatibility Functions

- $s(x, y) = \phi(x) \cdot \Theta(y)$

  Pros:
  - The representations are learned
  - Structure in label space can be ***discovered***
  - Inference can use fast (approx.) nearest neighbor lookup

  Cons:
  - The representations of images and labels need to lie in the same inner product space
  - Features need to ***correspond*** or ***align***

  Solution:
  - ***Learn to align*** representations

Compatibility Functions

- $s(x, y) = \phi^T(x)\, W\, \Theta(y)$

Compatibility Functions

$\otimes$ Outer product

vec(.) Converts mxn matrices to mn dim. vector

- $s(x,y) = \phi^T(x)\, W\, \Theta(y)$
  $= \mathrm{vec}(W) \cdot \mathrm{vec}(\phi(x) \otimes \Theta(y))$

Pros:

– Can ***learn to align*** representations

Cons:

– Relatively expensive to compute if m and n are large

– More parameters to learn ($m \times n$ parameters)

Solution:

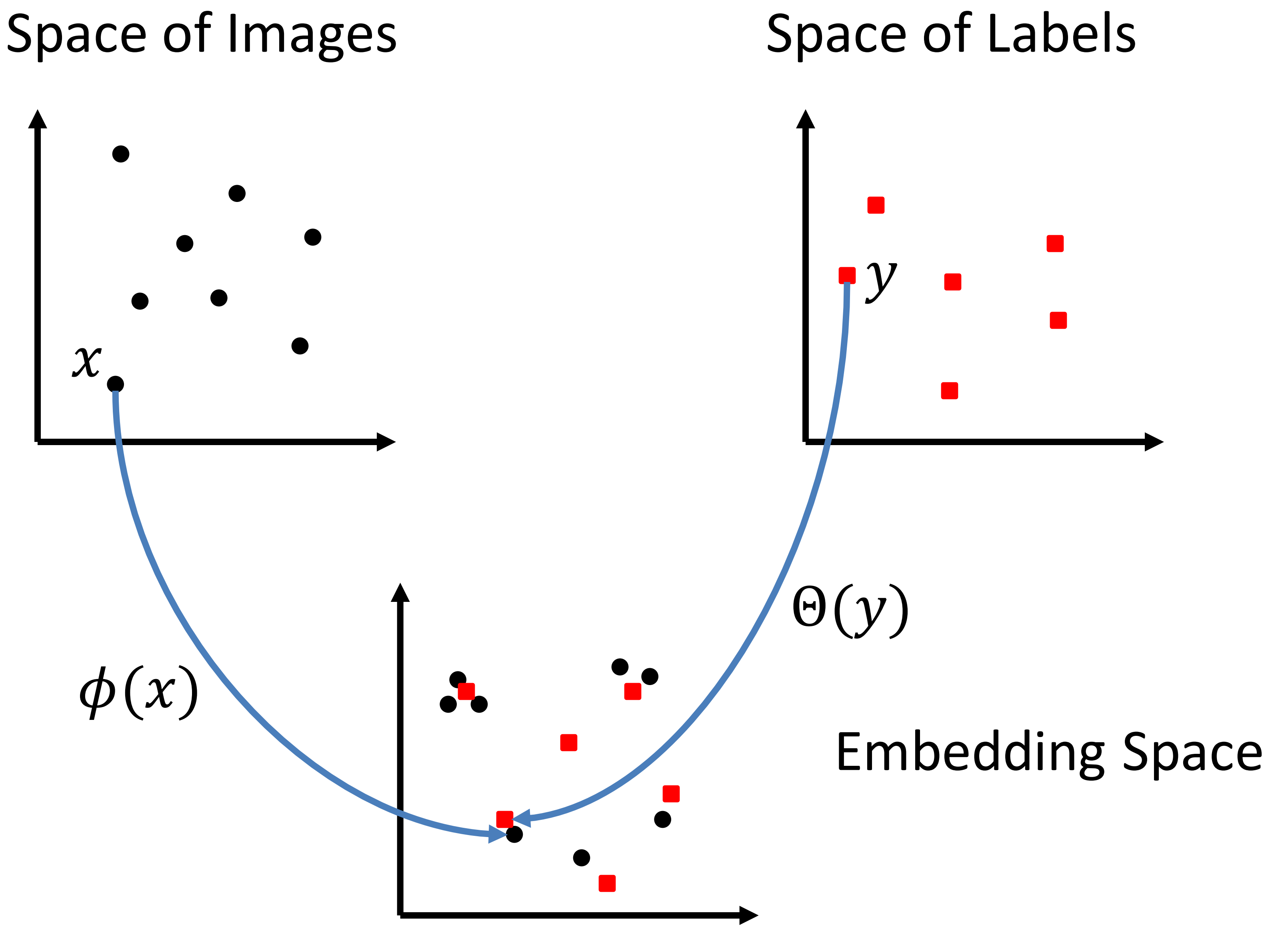
– Assume a ***low rank*** decomposition of W

– $W = U^T V$ where $U \in k \times m$ and $V \in k \times n$

– $s(x,y) = \phi^T(x)\, U^T V\, \Theta(y)$
  $= \big(U\phi(x)\big) \cdot \big(V\Theta(y)\big) = \phi'(x) \cdot \Theta'(y)$

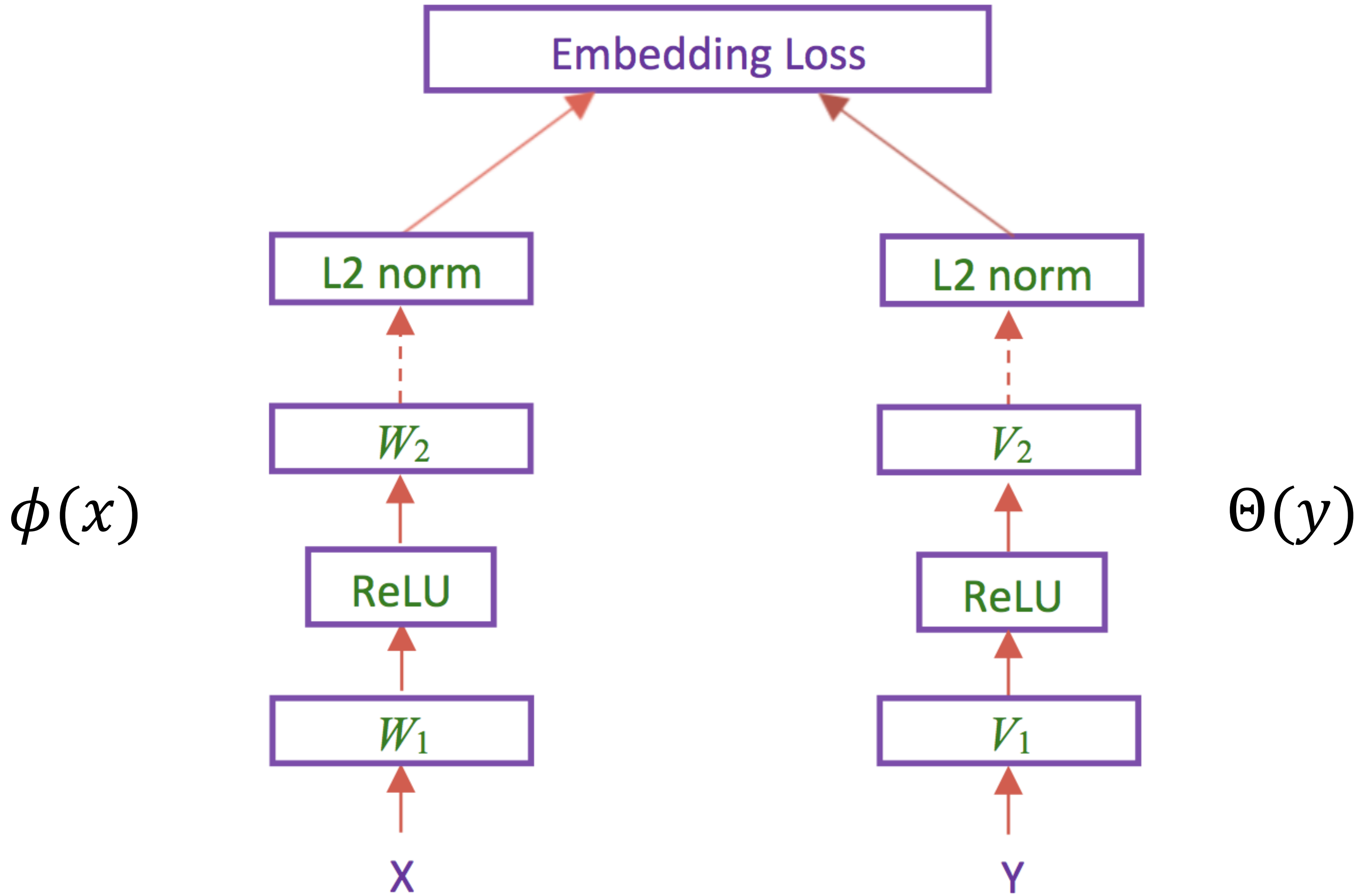
$k \times (m+n)$ parameters

Embeddings



Space of Images
Space of Labels
x
y
$\phi(x)$
$\Theta(y)$
Embedding Space

Embedding Network



Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning Two-Branch Neural Networks for Image-Text Matching Tasks." arXiv preprint arXiv:1704.03470 (2017).

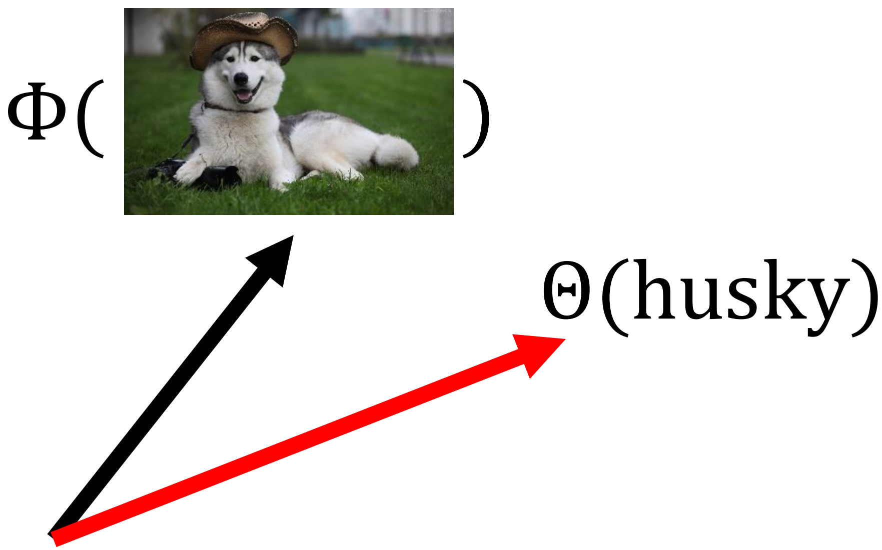
Embedding/Metric Learning

- Minimize ***distance*** between GT pairs
  - Ignores relative ranking
- Max-Margin Loss
  - Preserves ranking but O(#classes)
- Triplet Loss
  - Preserves ranking but O(constant)
- Bi-directional Ranking Loss
  - Good for bi-directional retrieval

Minimize distance between GT pairs

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

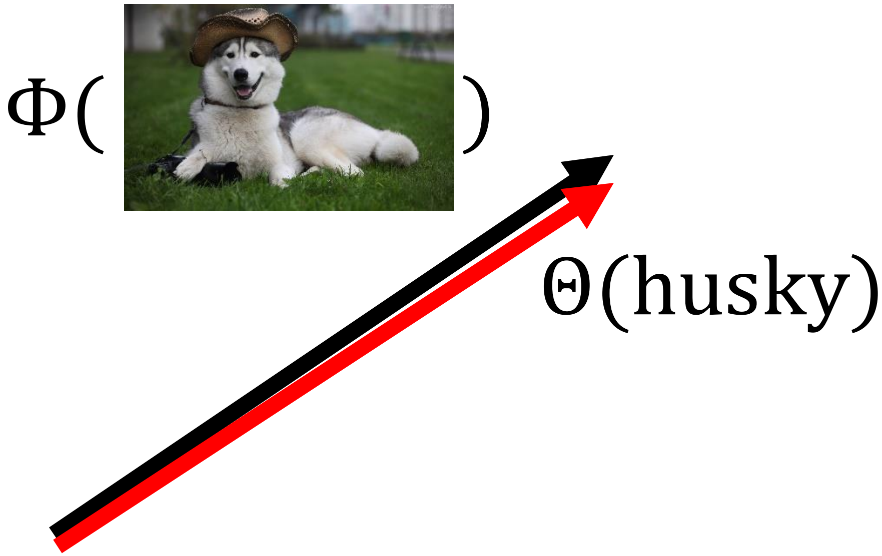
$$L(\phi, \Theta) = -\sum_i s(x, y; \phi, \Theta)$$



Minimize distance between GT pairs

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

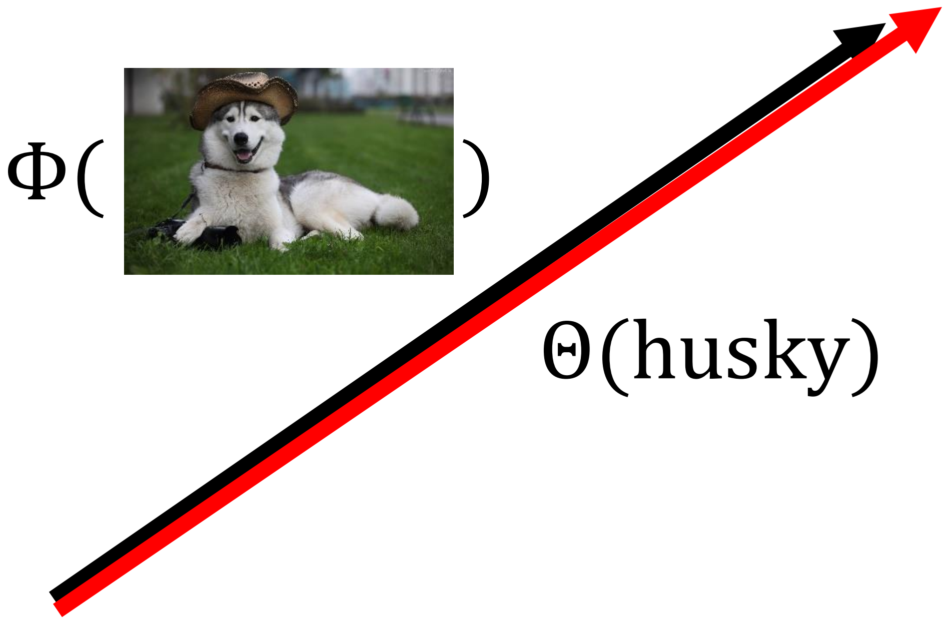
$$L(\phi, \Theta) = -\sum_i s(x, y; \phi, \Theta)$$



Minimize distance between GT pairs

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

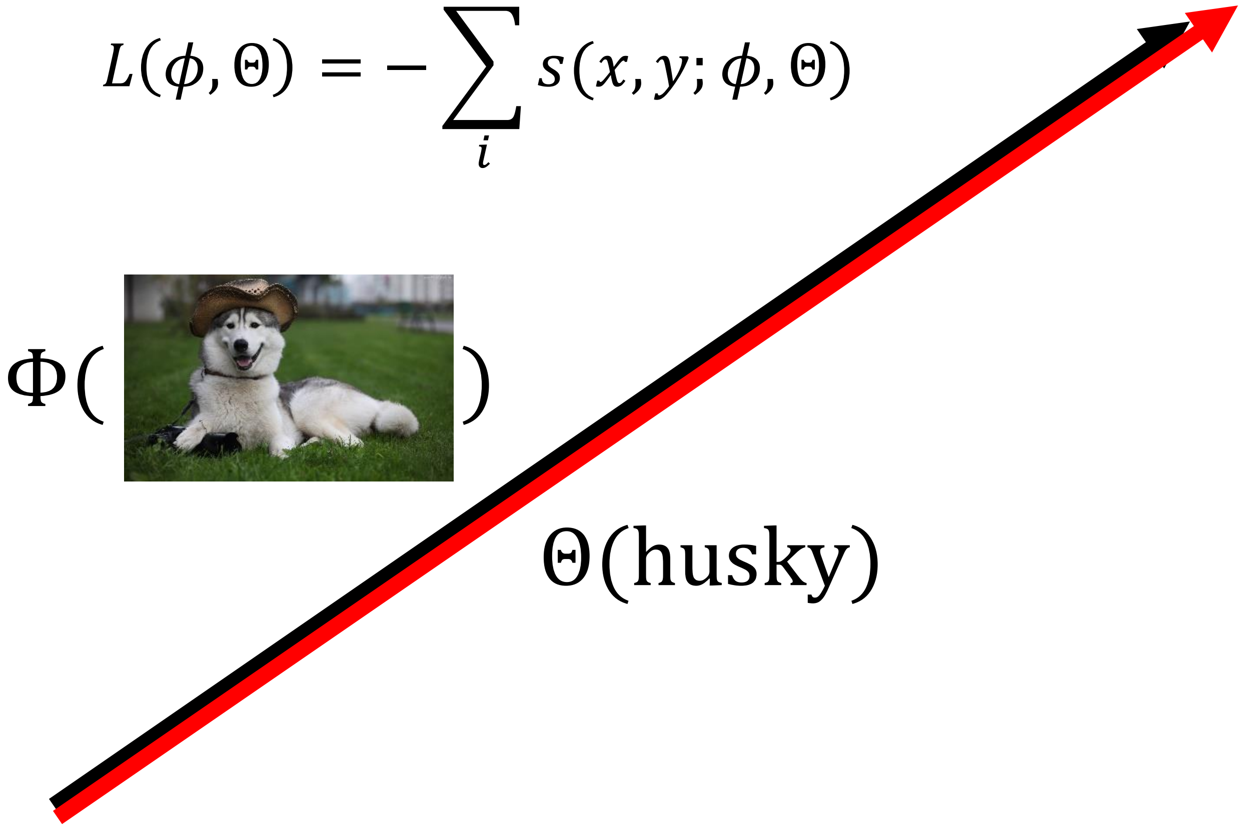
$$L(\phi, \Theta) = -\sum_{i} s(x, y; \phi, \Theta)$$



Minimize distance between GT pairs

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

$$L(\phi, \Theta) = -\sum_i s(x, y; \phi, \Theta)$$



Minimize distance between GT pairs

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

$$L(\phi, \Theta) = -\sum_i s(x, y; \phi, \Theta)$$

- Trivial Solution
  - Map all x and y to the same point at infinity for $\phi(x) \cdot \Theta(y)$
  - Ignores the relative score of labels for the same image!

- What we really want is for the correct label to have a high score while producing a lower score for incorrect labels

Max-Margin Loss

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

$$L(\phi, \Theta) = \sum_{i} \sum_{l \neq y_i} \max\{0, 1 + s(x_i, l) - s(x_i, y_i)\}$$

$\Phi($  $)^T \Theta(\text{husky}) \; > \; \Phi($  $)^T \Theta(\text{pug}) + 1$

$\Phi($  $)^T \Theta(\text{husky}) \; > \; \Phi($  $)^T \Theta(\text{poodle}) + 1$

⋮

Max-Margin Loss

- Ground truth image-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$

$$L(\phi, \Theta) = \sum_{i} \sum_{l \neq y_i} \max\{0, 1 + s(x_i, l) - s(x_i, y_i)\}$$

- Need to compute scores for all labels
  - Not scalable to large number of classes
  - Not suitable when multiple labels apply.
    - Eg. Dog and Husky

Triplet Loss

- Ground truth image-label triplets $\{(x_i, y_i^+, y_i^-)\}_{i=1}^N$

$$L(\phi, \Theta) = \sum_i \max\{0, 1 + s(x_i, y_i^-) - s(x_i, y_i^+)\}$$

$x_i =$  $\quad y_i^+ = \text{husky} \quad y_i^- = \text{human}$

Triplet Loss

- Ground truth image-label triplets $\{(x_i, y_i^+, y_i^-)\}_{i=1}^N$

$$L(\phi, \Theta) = \sum_i \max\{0, 1 + s(x_i, y_i^-) - s(x_i, y_i^+)\}$$

- Scalable and preserves label order ☺

- Common to provide more than one but small ($\ll$ #labels) number of negative example

Bi-direction Ranking Loss

- Ground truth image-label pairs $\{(x_i^+, x_i^-, y_i^+, y_i^-)\}_{i=1}^N$

$$L(\phi, \Theta) = \sum_i \max\{0, 1 + s(x_i^+, y_i^-) - s(x_i^+, y_i^+)\}$$
$$+ \sum_i \max\{0, 1 + s(x_i^-, y_i^+) - s(x_i^+, y_i^+)\}$$

$x_i^+ =$ 

$y_i^+ = \text{husky}$ $\quad y_i^- = \text{human}$

$x_i^- =$ 

Bi-direction Ranking Loss

- Ground truth image-label pairs $\{(x_i^+, x_i^-, y_i^+, y_i^-)\}_{i=1}^N$

$$L(\phi, \Theta) = \sum_i \max\{0, 1 + s(x_i^+, y_i^-) - s(x_i^+, y_i^+)\}$$
$$+ \sum_i \max\{0, 1 + s(x_i^-, y_i^+) - s(x_i^+, y_i^+)\}$$

- Useful when the goal is bi-direction retrieval
  - Image-Caption Retrieval

Canonical Correlation Analysis (CCA)

- A non deep-learning alternative
- Often provides a strong baseline for embedding approaches for image-caption retrieval
- For random vectors X and Y, finds projection matrices W and V which maximize the ***correlation*** between WX and VY.

Gong, Yunchao, et al. "Improving image-sentence embeddings using large weakly annotated photo collections." European Conference on Computer Vision. Springer International Publishing, 2014.

PART II

LANGUAGE REPRESENTATION

Representations

- **Image**
  – Outputs of CNNs, Spatial Pyramid, HOG …

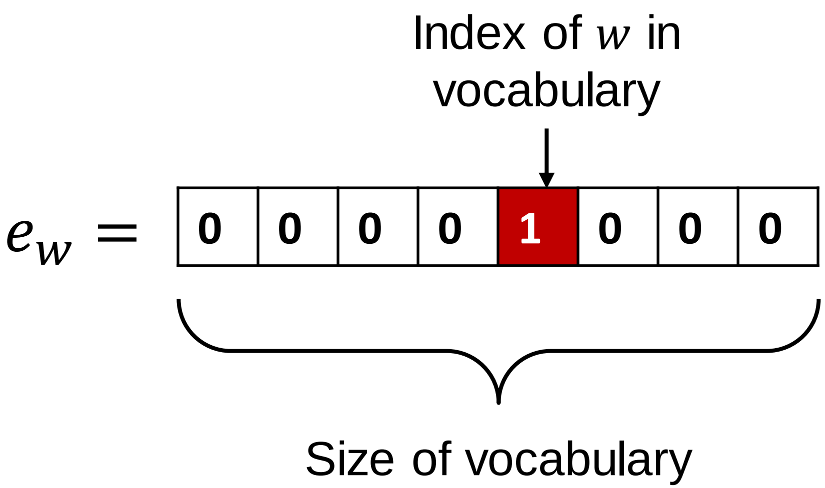
- **Language Labels**
  – Word:
    Object/Attribute recognition, scene classification

  – Phrase:
    Human-Object-Interaction, Visual Relationship Detection

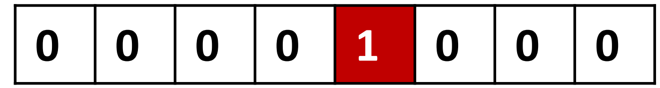
  – Sentence:
    Image Captioning, Visual Question Answering

Word Representations

- **One-Hot Encoding e_w**
  - Identity vector of the size of word vocabulary



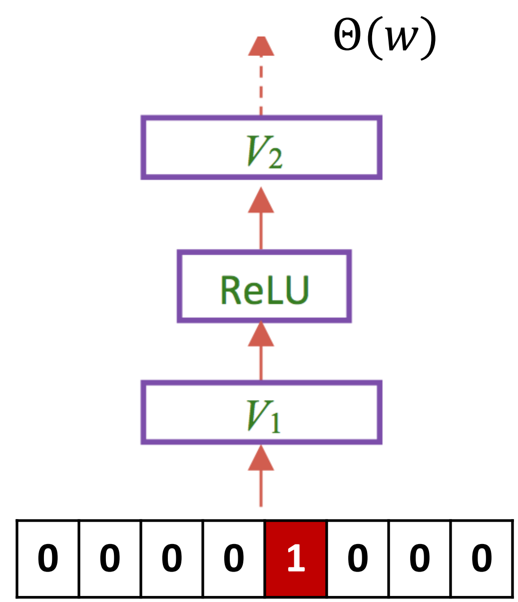
Word Representations



- **One-Hot Encoding e_w**
  - Identity vector of the size of word vocabulary

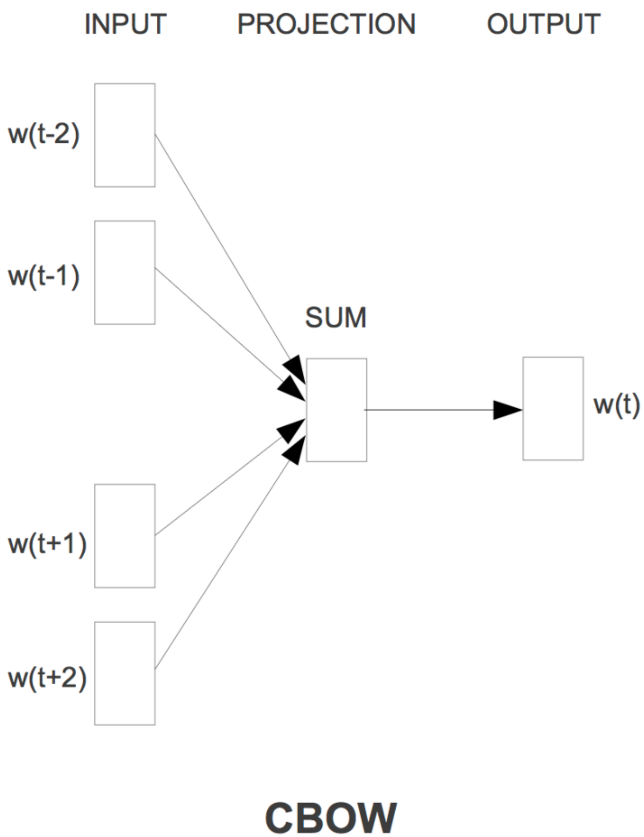
- **Linear/Non-Linear Transformation of e_w**
  - More compact representation than one-hot (300 vs 3M dim.)
  - Trained using large text corpuses (300B words, 3M vocab)
  - Capture ***semantics*** from training data
  - Eg. Word2vec, GloVe …

Textual data can be thought of as **external knowledge** in many vision tasks. Hence, common to learn Θ as a transformation of word2vec representation.



Word2Vec

- Words that occur in similar ***contexts*** are ***semantically*** similar

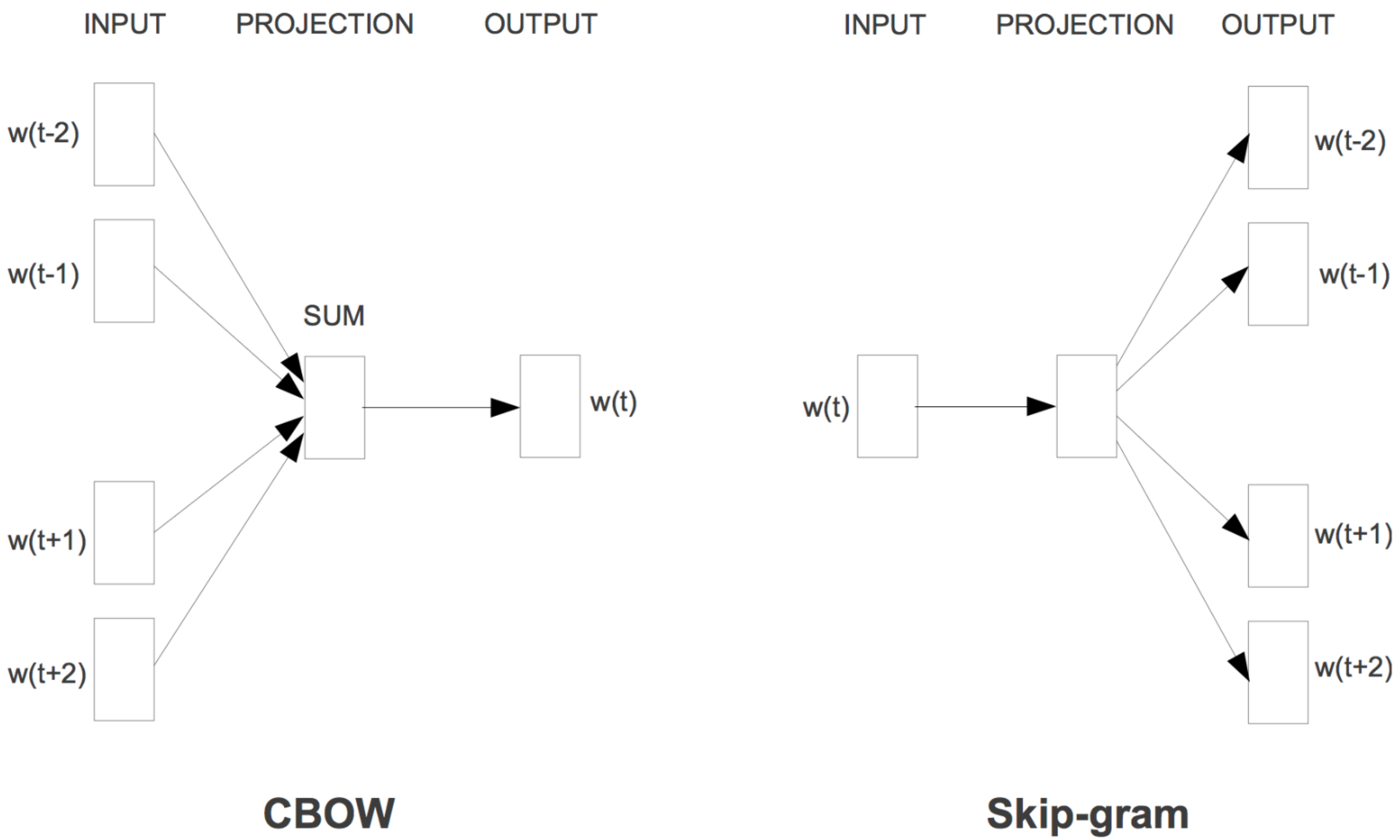


Predict word from context

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Word2Vec

- Words that occur in similar ***contexts*** are ***semantically*** similar



Predict word from context

Predict context from word

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Word2Vec Arithmetics

Japan – Sushi + Germany = ?

Word2Vec Arithmetics

Japan – Sushi + Germany = Bratwurst

bigger – big + small = ?

Word2Vec Arithmetics

Japan – Sushi + Germany = Bratwurst

bigger – big + small = smaller

Paris – France + Italy = ?

Word2Vec Arithmetics

Japan – Sushi + Germany = Bratwurst

bigger – big + small = smaller

Paris – France + Italy = Rome

similarity(tremendous, enormous) = 0.74

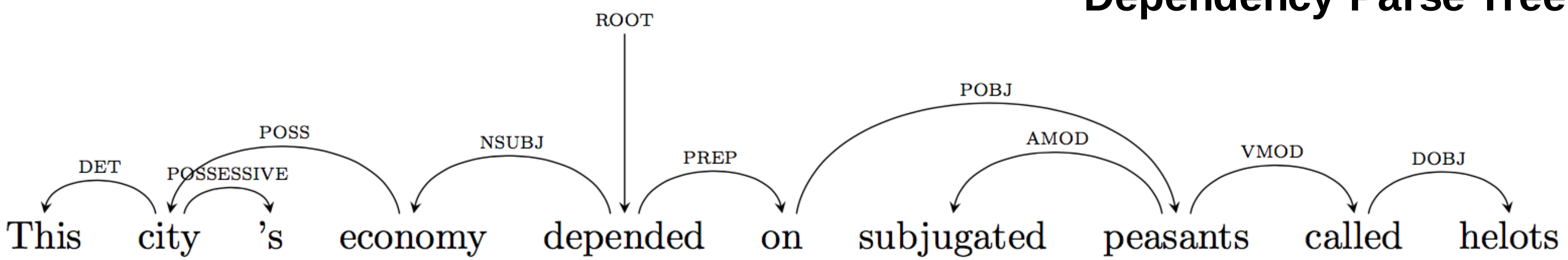
similarity(tremendous, negligible) = 0.37

most_similar(psyched) = geeked
excited
jazzed
bummed

https://quomodocumque.wordpress.com/2016/01/15/messing-around-with-word2vec/

Representing Phrases/Sentences

- Average or concatenate word representations
  - Simple
  - Works well for short and simple phrases “the brown cat”
- For complex sentences combine word representations guided by a parse tree



- Recurrent Neural Language Models

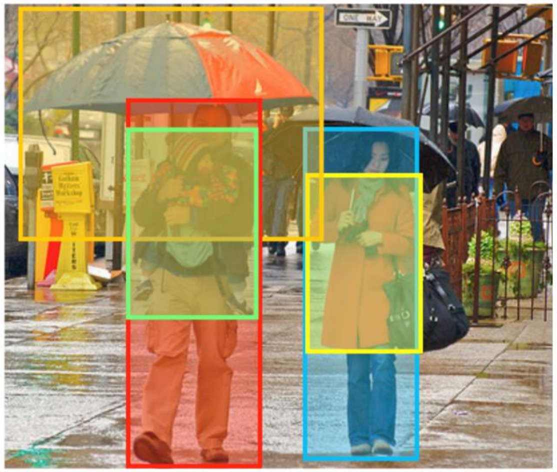
[Iyyer, Mohit, et al. "A Neural Network for Factoid Question Answering over Paragraphs." EMNLP. 2014.]

PART III

HOT TOPICS IN VISION-LANGUAGE RESEARCH

#TrendingInVisionLanguage

Phrase Localization



A man carries a baby under a red and blue umbrella next to a woman in a red jacket

Image Captioning

Caption: Man in black shirt playing a guitar.

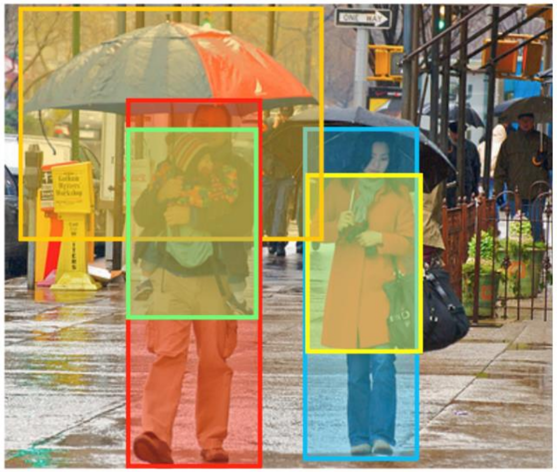
Visual Question Answering (VQA)

Question: What is the yellow object in the street?

Answer: Hydrant

#TrendingInVisionLanguage

Phrase Localization



A man carries a baby under a red and blue umbrella next to a woman in a red jacket

Image Captioning

Caption: Man in black shirt playing a guitar.

Visual Question Answering (VQA)

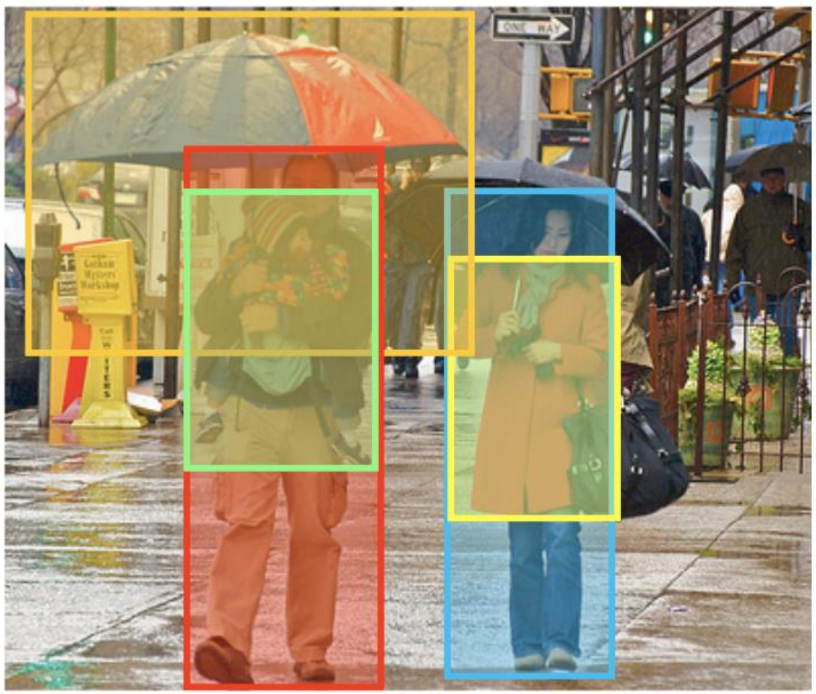
Question: What is the yellow object in the street?

Answer: Hydrant

Phrase Localization

Input Sentence and Image

A man carries **a baby** under **a red and blue umbrella** next to **a woman** in **a red jacket**



Cues	Examples
1) Entities	man, baby, umbrella, woman, jacket
2) Candidate Box Position	——
3) Candidate Box Size	——
4) Common Object Detectors	man → person baby → person woman → person
5) Adjectives	umbrella → red umbrella → blue jacket → red
6) Subject - Verb	(man, carries)
7) Verb – Object	(carries, baby)
8) Verbs	(man, carries, baby)
9) Prepositions	(baby, under, umbrella) (man, next to, woman)
10)Clothing & Body Parts	(woman, in, jacket)

Plummer, Bryan A., et al. "Phrase Localization and Visual Relationship Detection with Comprehensive Linguistic Cues." arXiv preprint arXiv:1611.06641 (2016).

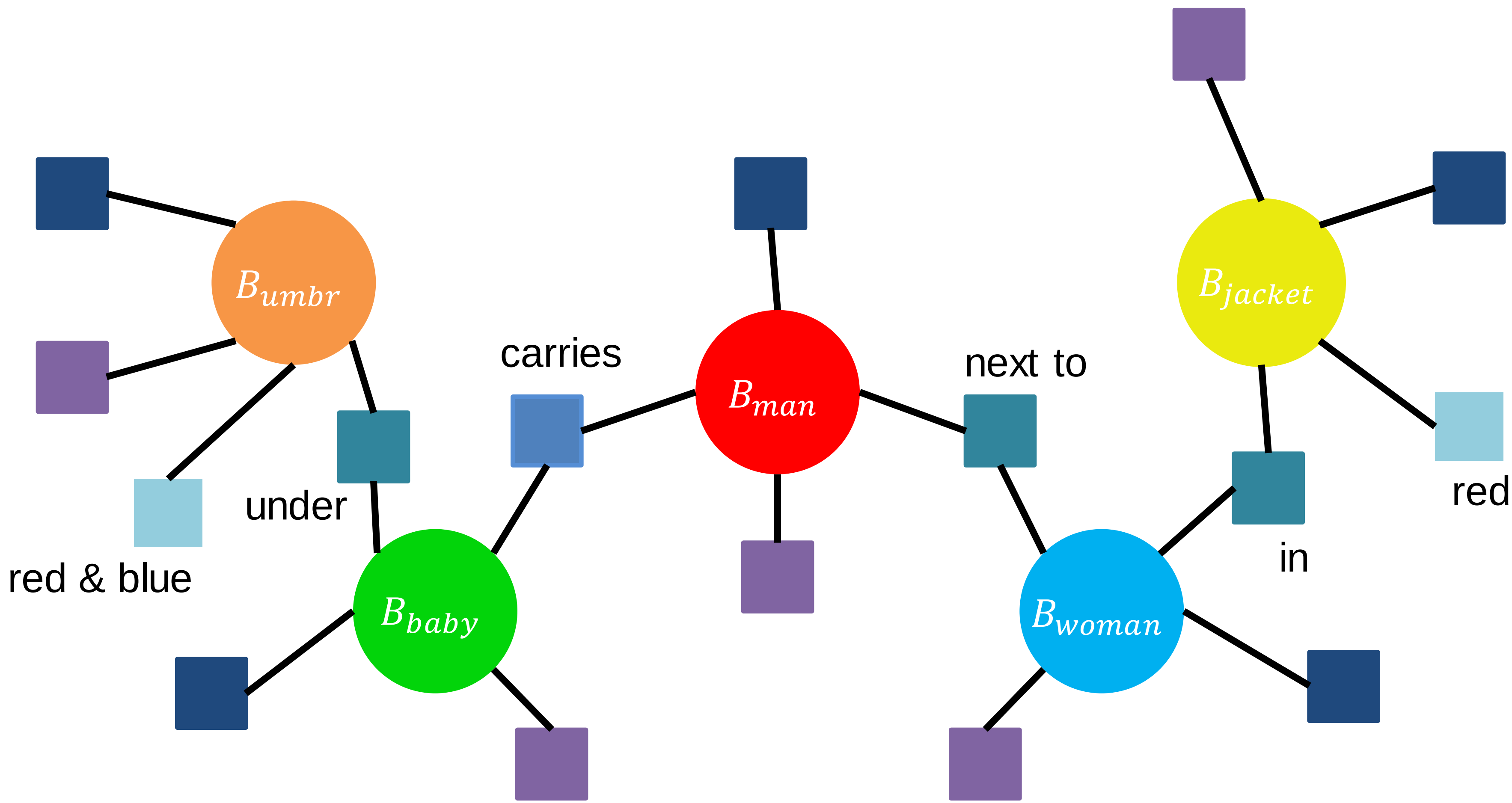
Appearance | Position & Shape | Adjective | Verb | Preposition

Input Sentence and Image

A man carries a baby under a red and blue umbrella next to a woman in a red jacket

Region Proposal Boxes $\{b_1, \cdots, b_n\}$

Find the most likely assignment of boxes to $(B_{umbr}, B_{man}, B_{baby}, B_{woman}, B_{jacket})$



Phrase Loc. As Energy Minimization

$$\min_{r_1,\cdots,r_N} \left\{ \sum_{p_i} S(p_i, r_i) \quad + \sum_{\rho_{ij}=(p_i, rel_{ij}, p_j)} Q(\rho_{ij}, r_i, r_j) \right\}$$

- Even 10 region proposals and 5 noun phrases lead to large search space (10^5)
- Fast inference methods
  - Graph Cuts with α-expansion
  - Belief Propagation (max-product)
  - Integer Quadratic Program Solvers

Plummer, Bryan A., et al. "Phrase Localization and Visual Relationship Detection with Comprehensive Linguistic Cues." arXiv preprint arXiv:1611.06641 (2016).

Learning factors for Phrase Loc.

- Factors
  - Given the ground truth bounding boxes
  - Each factor is trained separately
  - A mix of CCA based (appearance), SVM based (position) and hand coded factors (size)

- Weights
  - The relative weighting of the factors needs to be learned
  - Directly search for weights that maximize recall at IOU 0.5 (*fminsearch* in MATLAB)

Plummer, Bryan A., et al. "Phrase Localization and Visual Relationship Detection with Comprehensive Linguistic Cues." arXiv preprint arXiv:1611.06641 (2016).

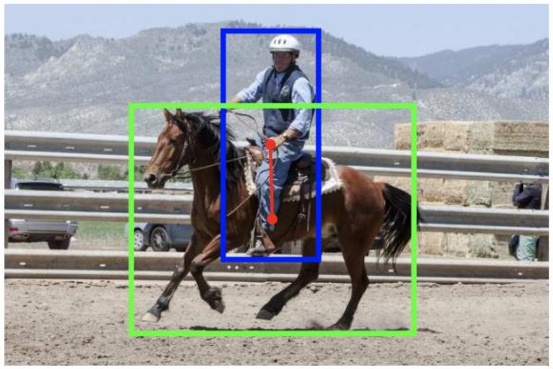
Effect of different cues on performance

	Method	Accuracy
(a)	**Single-phrase cues**	
	CCA	41.77
	CCA+Det	44.54
	CCA+Det+Size	49.77
	CCA+Det+Size+Adj	52.42
	CCA+Det+Size+Adj+Verbs	53.76
	CCA+Det+Size+Adj+Verbs+Pos (SPC)	**54.17**
(b)	**Phrase pair cues**	
	SPC+Verbs	54.23
	SPC+Verbs+Preps	54.34
	SPC+Verbs+Preps+C&BP (SPC+PPC)	**54.88**
(c)	**State of the art**	
	SMPL [33]	42.08
	NonlinearSP [32]	43.89
	GroundeR [26]	47.81
	MCB [5]	48.69
	RtP [25]	50.89

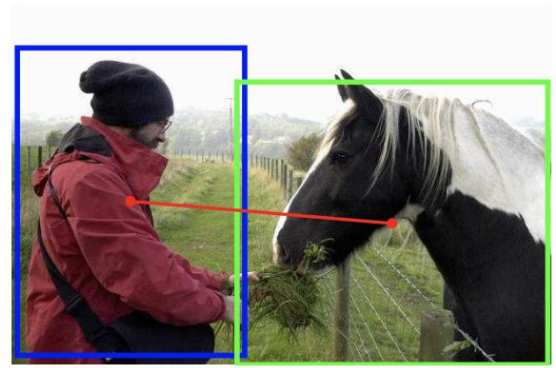
Results on Flickr 30k Entities

Similar problems

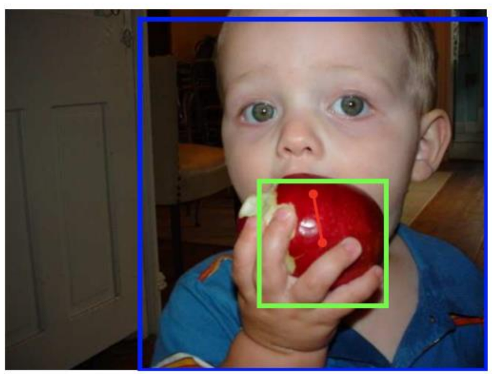
- Human-Object Interaction



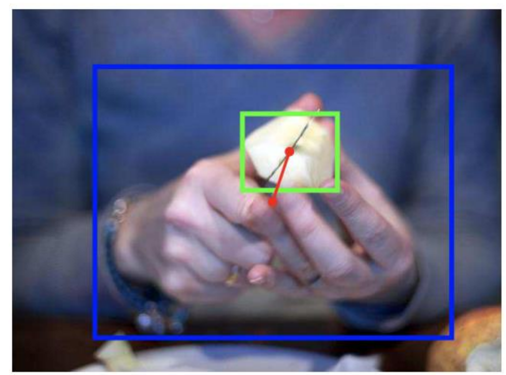
(a) riding a horse



(b) feeding horses



(c) eating an apple

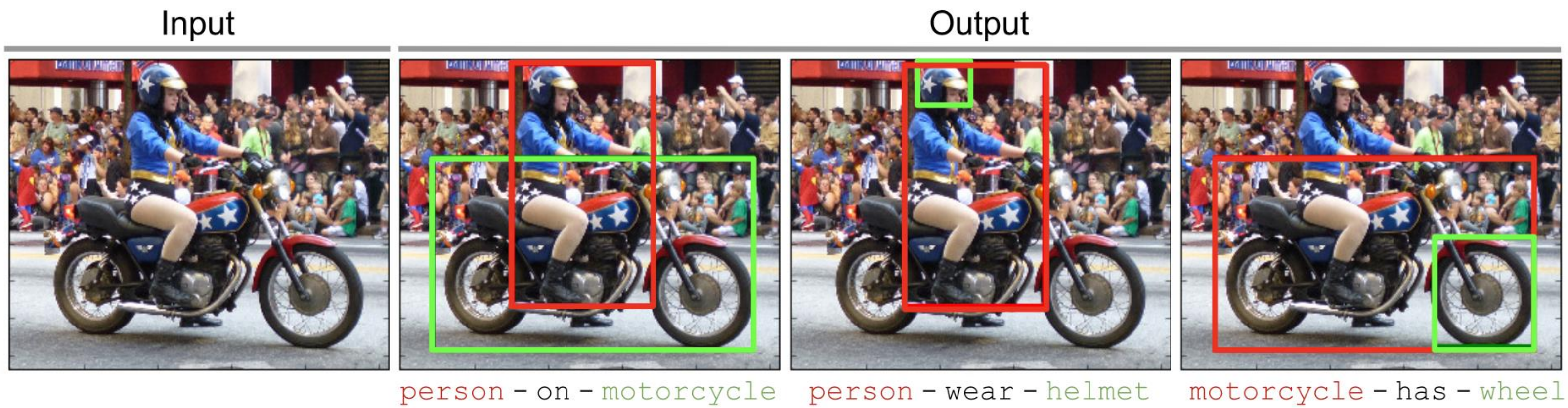


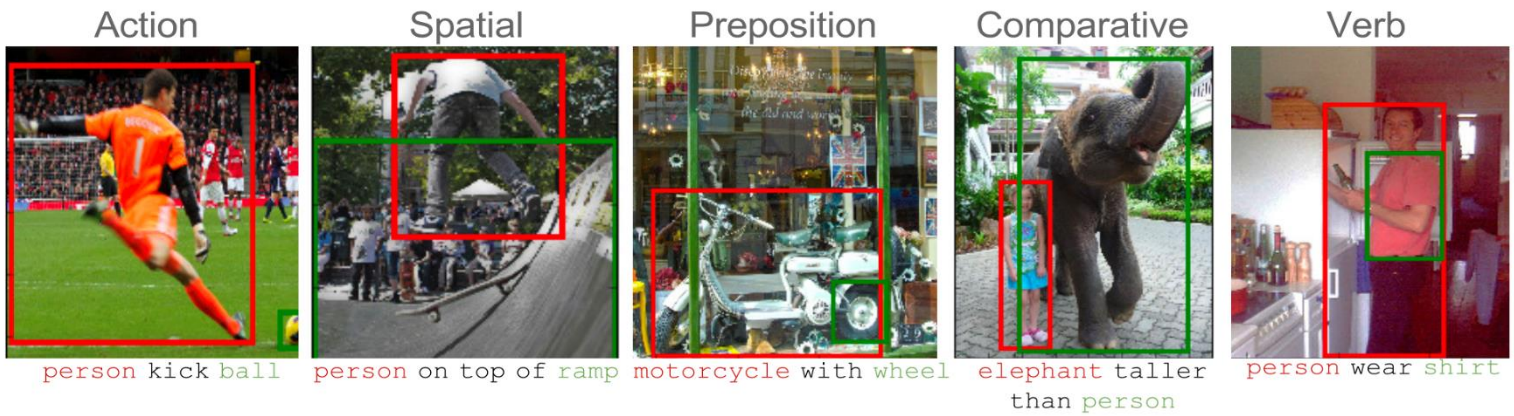
(d) cutting an apple

Chao, Yu-Wei, et al. "Learning to Detect Human-Object Interactions." arXiv preprint arXiv:1702.05448 (2017).

Similar problems

- Visual relationship detection
  - Object-Object Interaction as well





Lu, Cewu, et al. "Visual relationship detection with language priors." European Conference on Computer Vision. Springer International Publishing, 2016.

Similar problems

- Referring Expression Comprehension
  - Multiple confusable objects

A girl wearing glasses and a pink shirt.

An Asian girl with a pink shirt eating at the table.

A boy brushing his hair while looking at his reflection.

A young male child in pajamas shaking around a hairbrush in the mirror.

A woman in a flowered shirt.

Woman in red shirt.

The woman in black dress.

A lady in a black dress cuts a wedding cake with her new husband.

Mao, Junhua, et al. "Generation and comprehension of unambiguous object descriptions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Similar problems

- Visual Semantic Role Labelling / Situation Recognition
  - Appearance changes with actions (clipping vs jumping)
  - Different ***situation*** with same action can look different

CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

Yatskar, Mark, Luke Zettlemoyer, and Ali Farhadi. "Situation recognition: Visual semantic role labeling for image understanding." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016

Open questions

- How are these tasks related?
  - What is common among them?
  - What is different?
- Is there a single computational model for them?
- Good and extensible representations?
- Connections to weakly supervised learning?

#TrendingInVisionLanguage

Phrase Localization

A man carries a baby under a red and blue umbrella next to a woman in a red jacket

Image Captioning

Caption: Man in black shirt playing a guitar.

Visual Question Answering (VQA)

Question: What is the yellow object in the street?

Answer: Hydrant

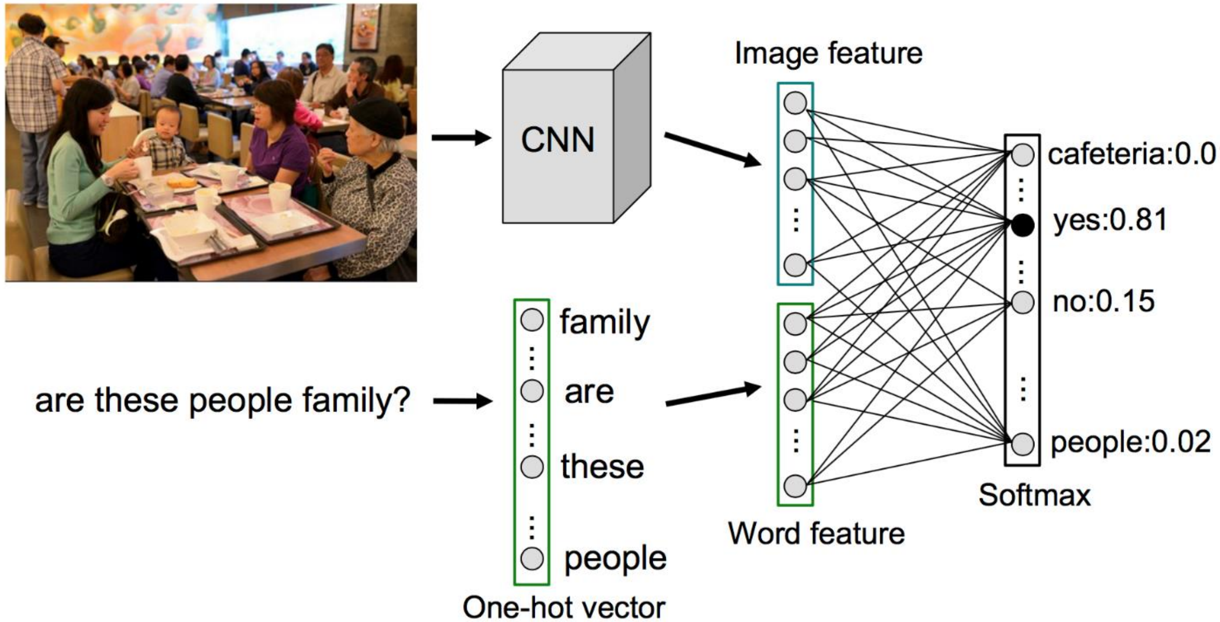
Visual Question Answering

Question:
Are these people family?

Answer:
Yes

- What do “people” look like?
- What makes a group of people “family”?
- Understand the question
- Verify answer
  - Is the answer valid for the given question (language prior)
  - Does the answer apply to the image (visual verification)

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).]

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)

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

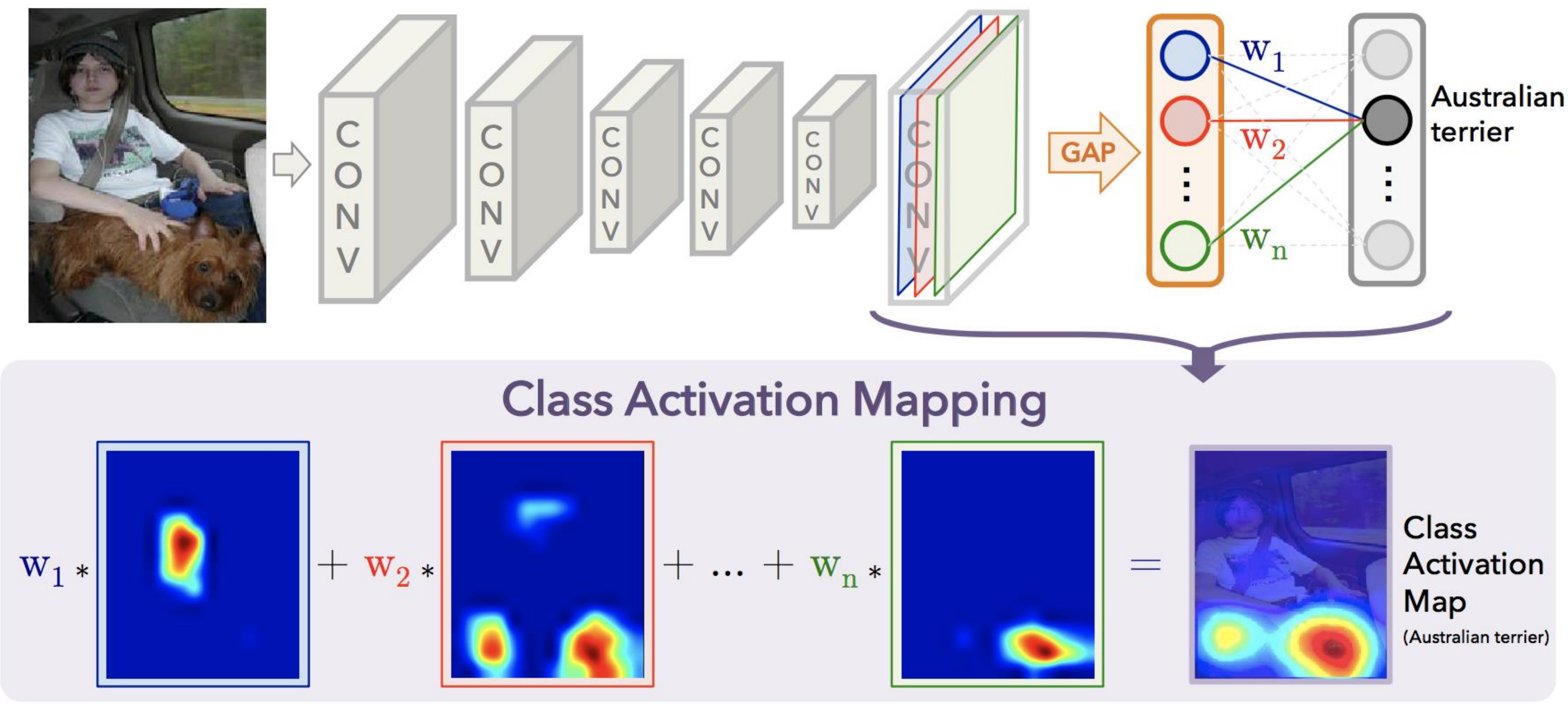
Quantitative Evaluation

	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

Does the model learn to localize?

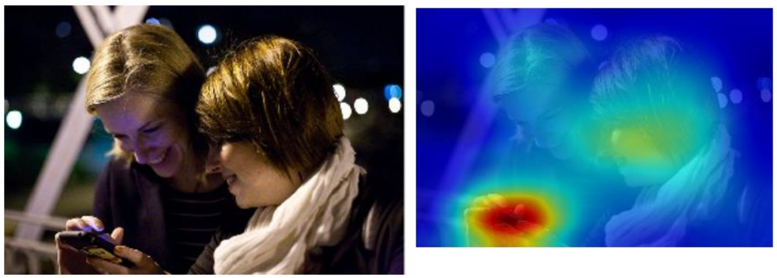
- Class Activation Mapping
  - Technique to generate localization heat maps from classification networks



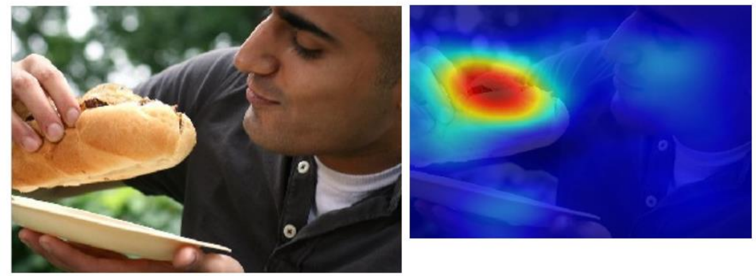
[Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.]

Does the model learn to localize?

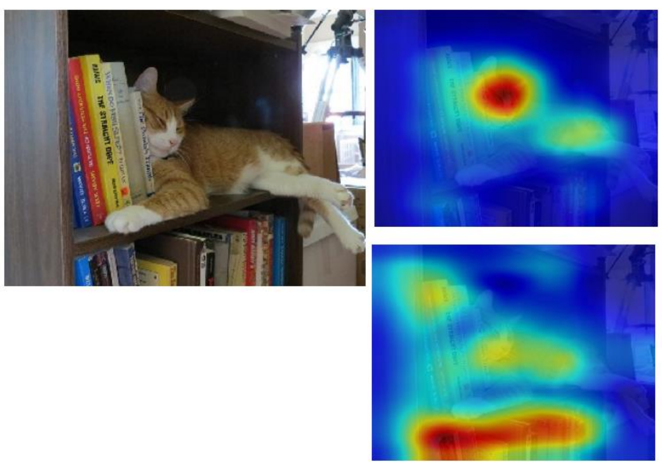
- Class Activation Mapping applied to VQA Baseline



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)

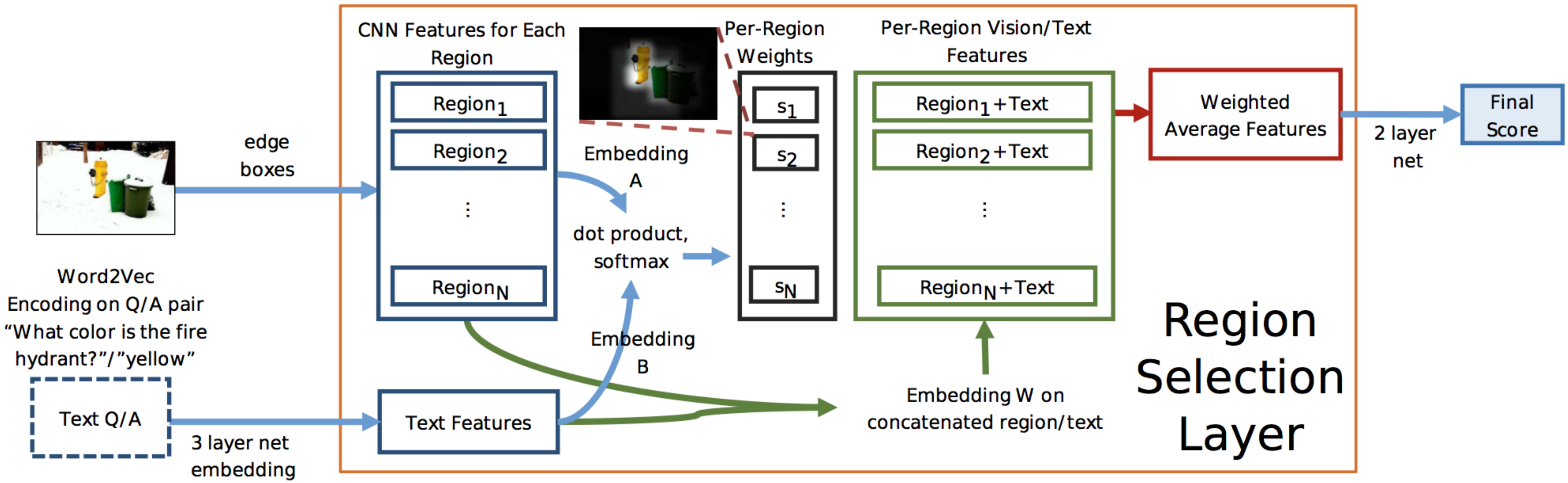
VQA models with explicit localization

Neural models with ***attention***

Treat the ***relevance*** of each location (pixels/region proposals) as ***latent*** variables

- Encodes our intuition
  - Need to ***look*** at the right region to answer the question
  - Need to look at the hat to answer "What color is the person's hat?"
- Possibly reduces ***model complexity***
  - Bias-Variance Tradeoff
- Improves ***interpretability***

VQA Model that knows "Where To Look"



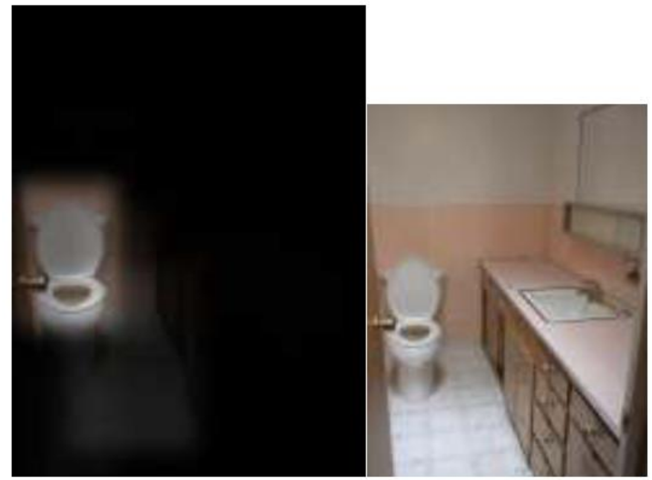
[Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.]()

VQA Model that knows "Where To Look"

What color on the stop light is lit up?

L: red (-0.1)
I: red (-0.8)
R: green (1.1)
Ans: green

What room is this?



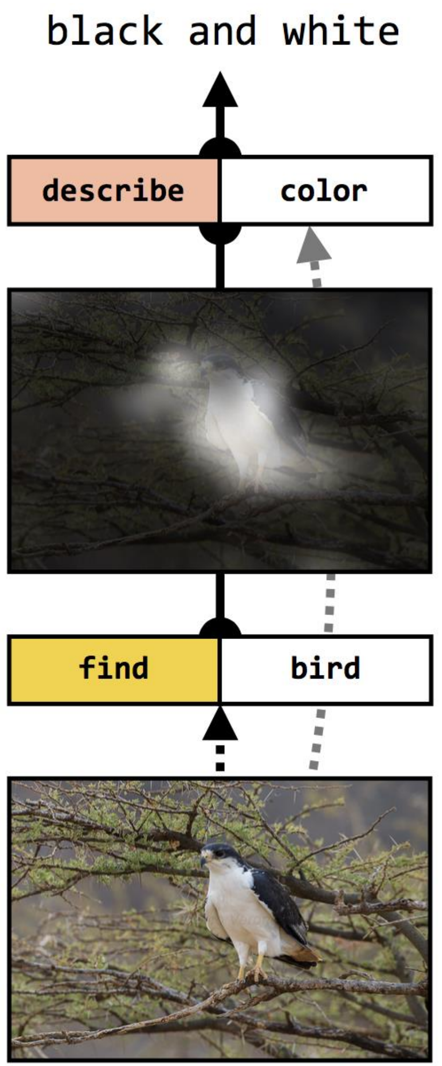
L: bathroom(0.1)
I: bathroom (2.6)
R: bathroom (6.8)
Ans: bathroom

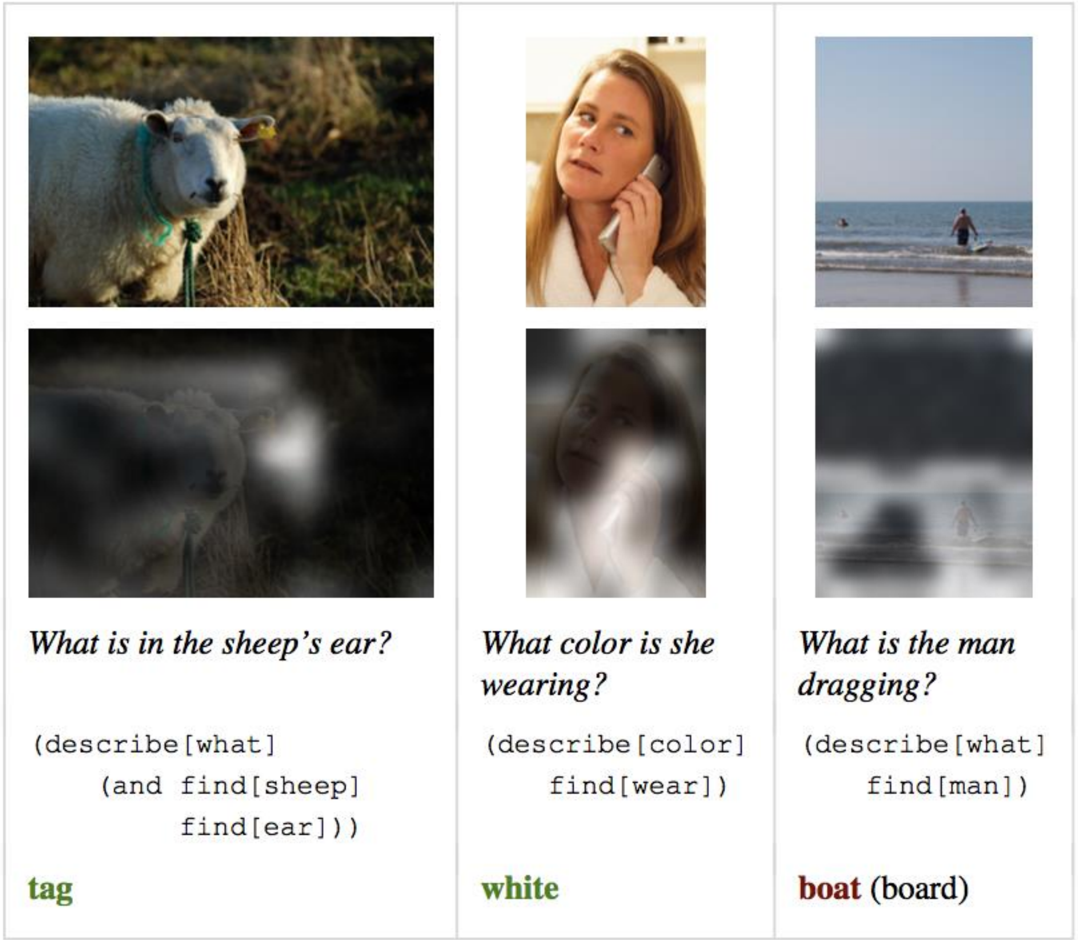
Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

A shift towards *compositional* models

- Answering a question can be divided into subtasks
- Design ***components/modules*** for each subtask
- Given a question
  - Decide which modules to use and the arrangement of modules ***on the fly*** based on the question parse tree
  - Execute the constructed compositional model on the image

Neural Module Networks





Q: What color is the bird?

Andreas, Jacob, et al. "Learning to compose neural networks for question answering." arXiv preprint arXiv:1601.01705 (2016).

#TrendingInVisionLanguage

Phrase Localization

A man carries a baby under a red and blue umbrella next to a woman in a red jacket

Image Captioning

Caption: Man in black shirt playing a guitar.

Visual Question Answering (VQA)

Question: What is the yellow object in the street?

Answer: Hydrant

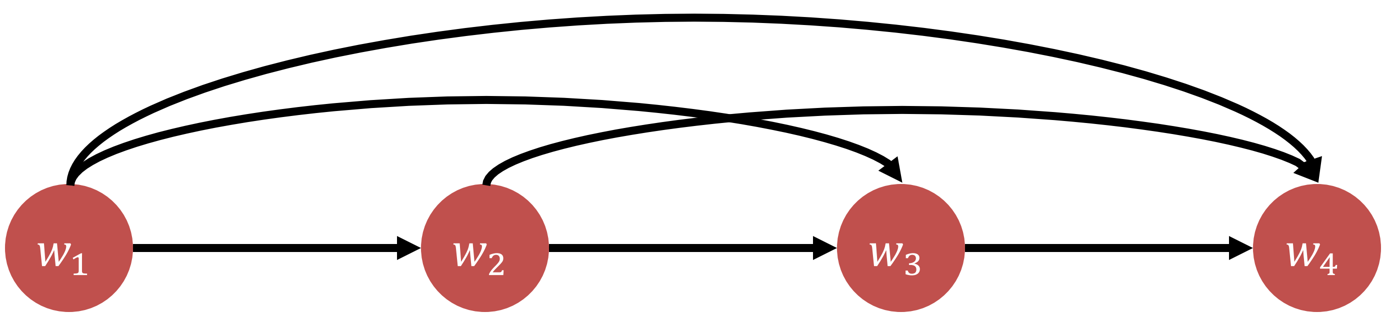
Representing Phrases/Sentences

- Using recurrent neural language models

Word-level Language Model

Given the sequence of words $w_1, \cdots, w_n$ that form a sentence, produces likelihood of the sentence

$$P(w_1, \cdots, w_n) = \prod_i P(w_i | w_{1:i-1})$$

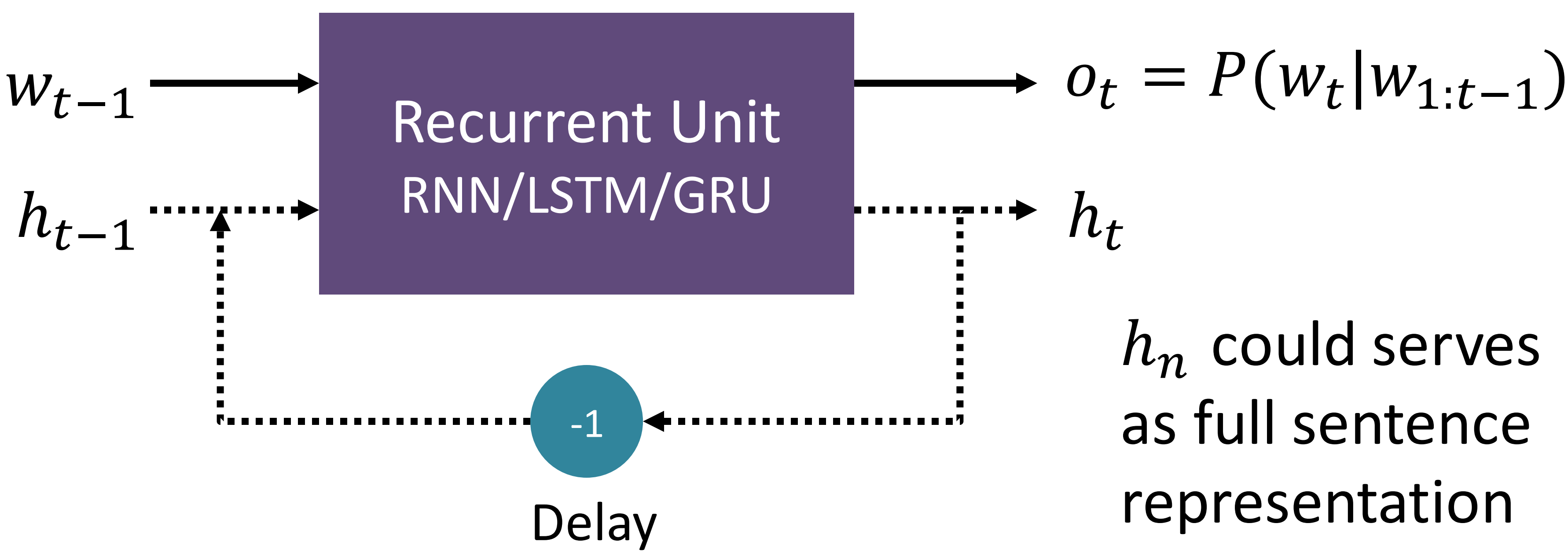


Representing Phrases/Sentences

- Using recurrent neural language models

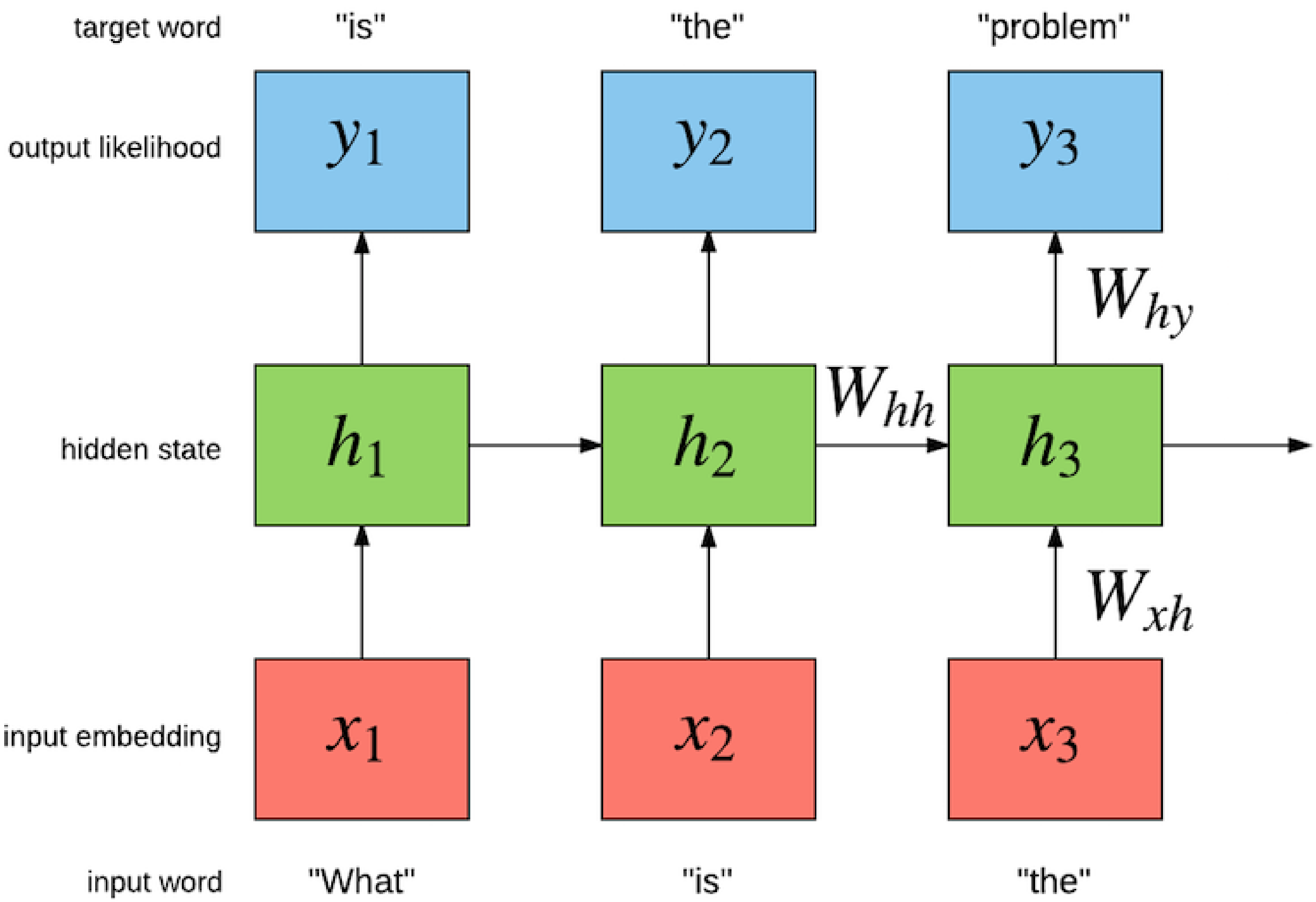
Recurrent Models

$$P(w_1, \cdots, w_n) = \prod_i P(w_i | w_{1:i-1})$$



Representing Phrases/Sentences

- Recurrent neural language model ***unfolded in time***



Representation of phrase "What is the"

Training Recurrent Language Models

- Trained on domain specific text data
  - News articles, image captions, Linux code base …
- Parameters learned through maximization of likelihood of ground truth text using ***Back-Propagation Through Time (BPTT)***
- Gating mechanism used to overcome ***vanishing/exploding gradients*** due to chain rule

Image Captioning

- Generate caption given image

- Training involves learning

$$P(S|I;\theta) = P(s_1, \cdots, s_n|I;\theta)$$

- Generation involves sampling from $P(S|I)$ or performing MAP inference

$$S^* = \text{argmax}\, P(S|I)$$

Image Captioning Model

$$P(S|I;\theta) = P(s_1, \cdots, s_n|I;\theta)$$

- Recall we modelled $P(s_1, \cdots, s_n)$ for language models

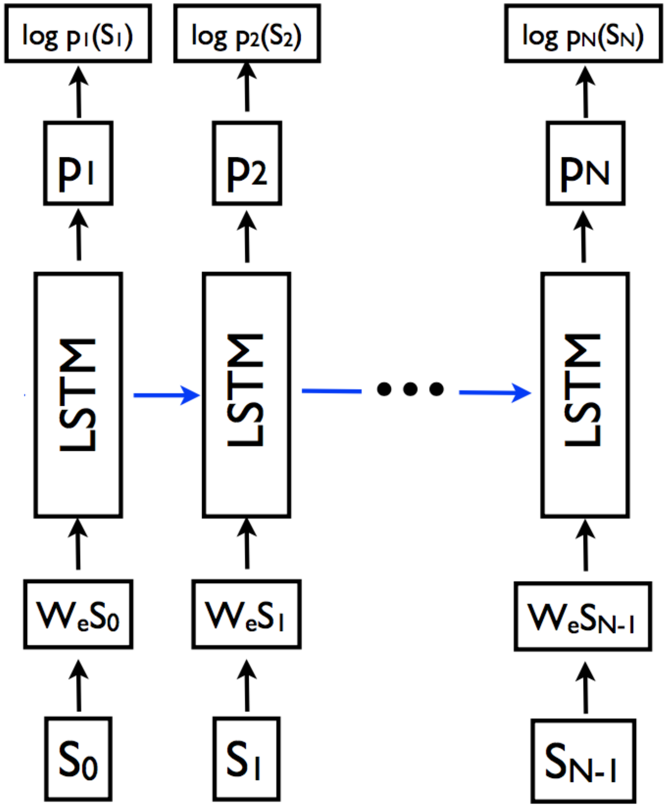
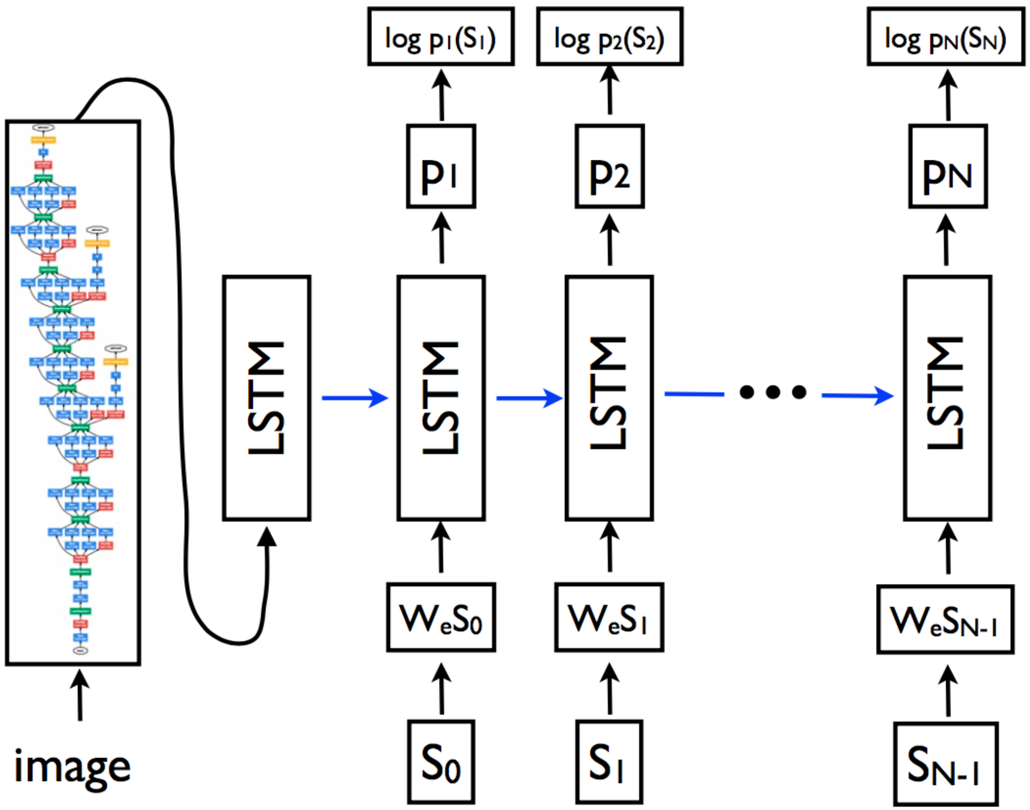


Image Captioning Model

$$P(S|I;\theta) = P(s_1, \cdots, s_n|I;\theta)$$

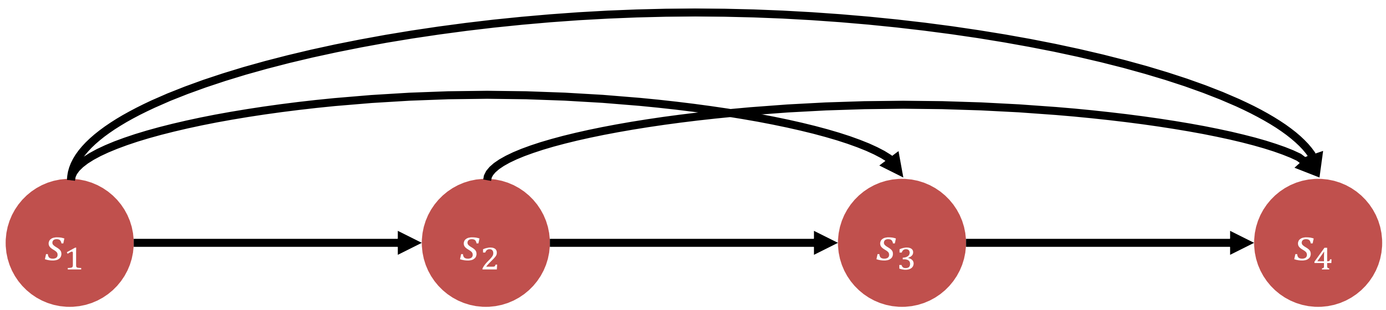
- Recall we modelled $P(s_1, \cdots, s_n)$ for language models



Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

Image Caption Generation

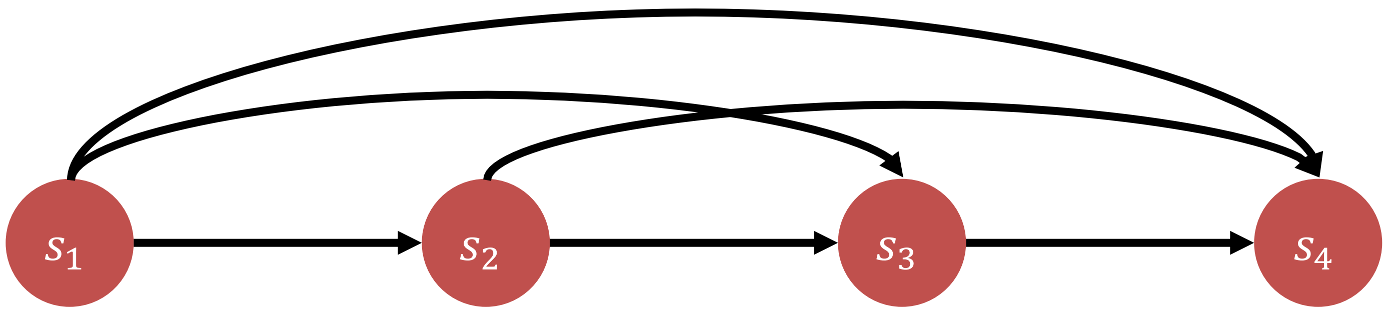
- How to sample from $P(S|I)$



$$s_1 \sim P(s_1|s_o, I)$$
$$s_2 \sim P(s_2|s_{o:1}, I)$$
$$s_3 \sim P(s_3|s_{o:2}, I)$$
$$s_4 \sim P(s_4|s_{o:3}, I)$$

Image Caption Generation

- MAP inference on $P(S|I)$
- Beam Search for approximate inference



$$5 \times s_1 \sim P(s_1 | s_o, I)$$

$$25 \times (s_1, s_2) \sim P(s_2 | s_{o:1}, I) \text{ (keep top 5)}$$

$$25 \times (s_1, s_2, s_3) \sim P(s_3 | s_{o:2}, I) \text{ (keep top 5)}$$

$$25 \times (s_1, s_2, s_3, s_4) \sim P(s_4 | s_{o:3}, I) \text{ (keep top 1)}$$

Qualitative Results

Evaluation

- Bleu Score
  - Given a candidate (machine generated) caption
  - Compare to reference (human annotated) captions
  - ***Modified n-grams*** word precision

Candidate: `the the the the the the the.`

Reference 1: `The cat is on the mat.`

Reference 2: `There is a cat on the mat.`

Uni-gram Precision = 7/7 **BLEU-1** = 2/7 **BLEU-2** = 0/6

  - Prefers shorter captions

    Eg. “`the cat`” has **BLEU-2** = 2/2

Key Takeaways

- Advantages of embedding based recognition
  - Scalability
  - Structure in label space
  - Use external knowledge
- Ways of representing words/phrases/sentences
  - Use context : Word2vec
  - Use language model : RNN/LSTM
- Vision-Language Applications:
  - Using multiple cues improves localization
  - Attention mechanisms make models more interpretable
  - Image Captioning models combine classification networks with language models but tricky to evaluate