Vision and Language

CS60010: Deep Learning

Abir Das

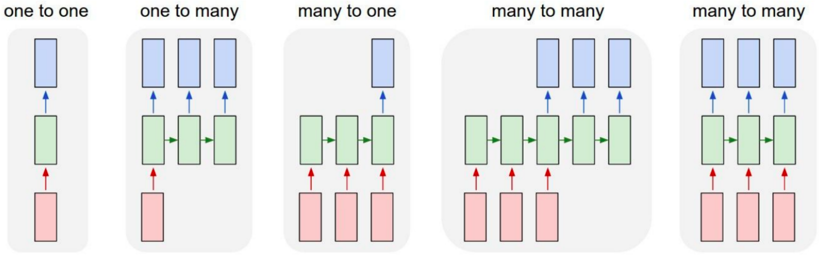
IIT Kharagpur

March 19 and 26, 2020

Agenda

To have a brief introduction to tasks that involve both computer vision and natural language processing and to get familiar with some approaches to address them.

Recap



Recap

- We've seen CNN based computer vision systems are fairly good at answering "what" (classification) and "where" (localization/ detection) by predicting a fixed of objects or scenes.
- In human analogy, this is much like a baby/toddler.

baby

toddler

Recap

- We've seen CNN based computer vision systems are fairly good at answering "what" (classification) and "where" (localization/ detection) by predicting a fixed of objects or scenes.
- In human analogy, this is much like a baby/toddler.


baby


toddler


preschooler

- In contrast, by age 4-5 a preschooler can look at a sequence of pictures and describe the depicted events in detailed sentences, as well as answer complex questions including "when", "how?", "how many?" (up to 10) "which?" *etc*.
- To advance the state of the art toward the preschooler level, we need models that integrate vision and language.

How can we Connect Vision and Language

- **Image description or image captioning**
  - A crowd of people looking at giraffes in a zoo.

How can we Connect Vision and Language

- **Image description or image captioning**
  - A crowd of people looking at giraffes in a zoo.
- **Referring expressions**
  - Person taking a photo.

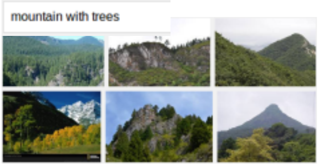
How can we Connect Vision and Language

- **Image description or image captioning**
  - A crowd of people looking at giraffes in a zoo.
- **Referring expressions**
  - Person taking a photo.
- **Question answering**
  - What time of year is it? - Ans: summer.

CS591 course, Boston University

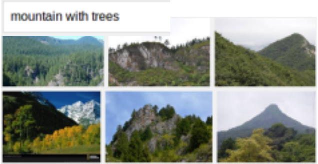
Applications

Image and video retrieval by content



Applications

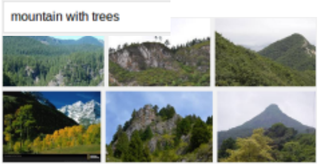
Image and video retrieval by content



Video description service

Applications

Image and video retrieval by content



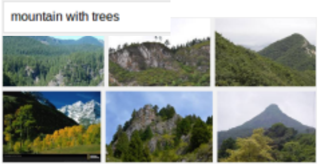
Video description service

Human Robot Interaction

CS591 course, Boston University

Applications

Image and video retrieval by content



Video description service

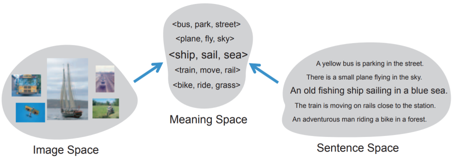
Human Robot Interaction

Video surveillance

CS591 course, Boston University

Image Captioning - Before Deep Learning

- Many early works on Image Description Farhadi et. al. ECCV10, Kulkarni et. al. CVPR11 identify objects, actions, attributes, and combine with linguistic knowledge to “tell a story”.
- Learn object, action, scene classifiers
- Estimate most likely agents and actions.
- Use template to generate sentence.



Farhadi et. al. ECCV10

Yu et. al. ACL’13

Image Captioning - Before Deep Learning

§ Many early works on Image Description Farhadi et. al. ECCV10, Kulkarni et. al. CVPR11 identify objects, actions, attributes, and combine with linguistic knowledge to “tell a story”.

§ Learn object, action, scene classifiers

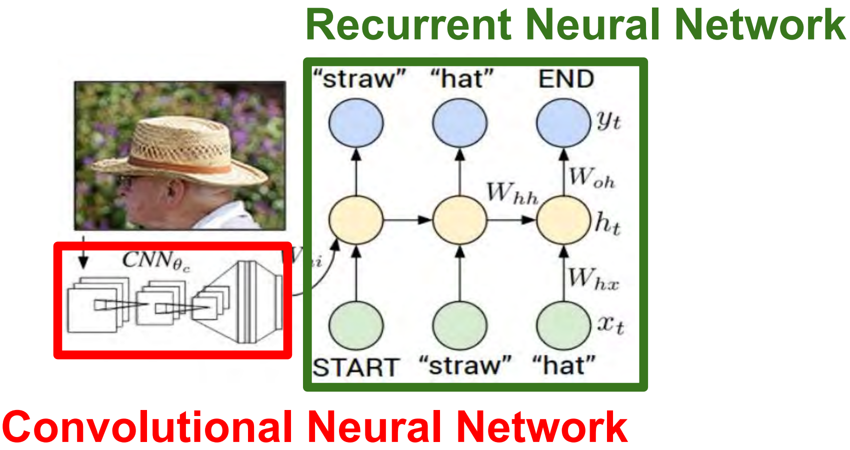
§ Estimate most likely agents and actions.

§ Use template to generate sentence.

§ Limitations

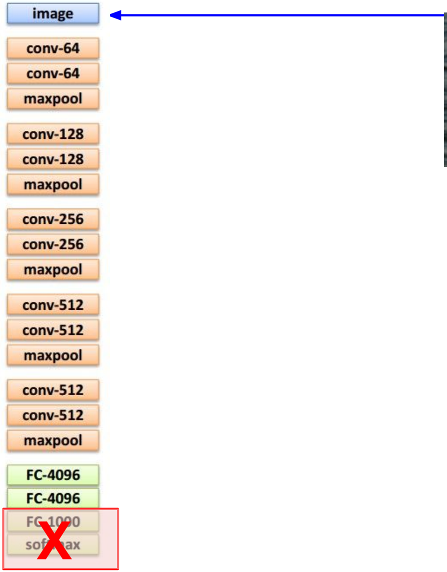
- Narrow Domains
- Small Grammars
- Template based sentences
- Mostly hand designed features

Image Captioning



CS231n course, Stanford University

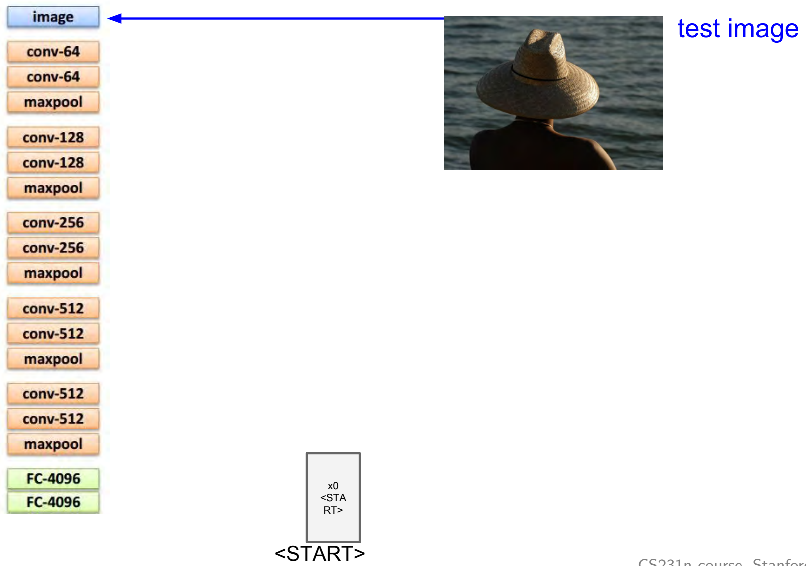
Image Captioning



test image

CS231n course, Stanford University

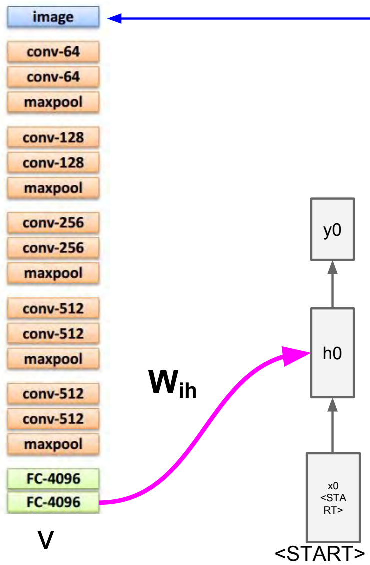
Image Captioning



CS231n course, Stanford University

Image Captioning


test image



before (No Image):

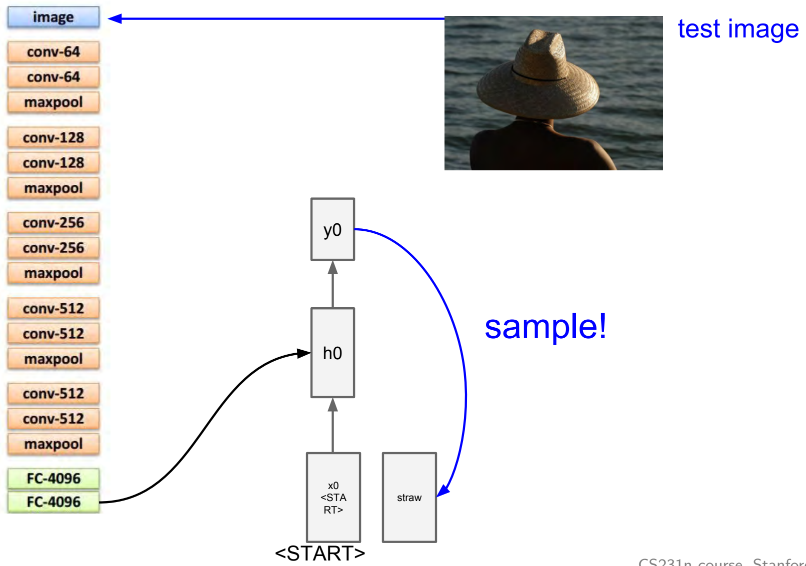
$h = \tanh(W_{xh} * x + W_{hh} * h)$

now (With Image):

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

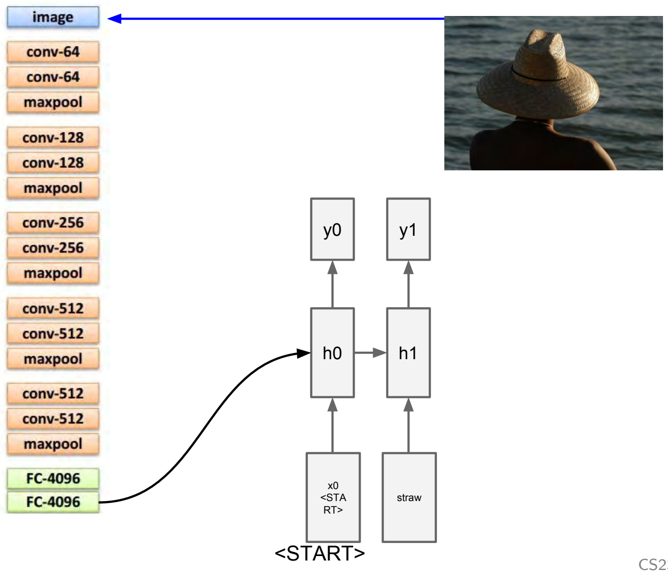
CS231n course, Stanford University

Image Captioning



CS231n course, Stanford University

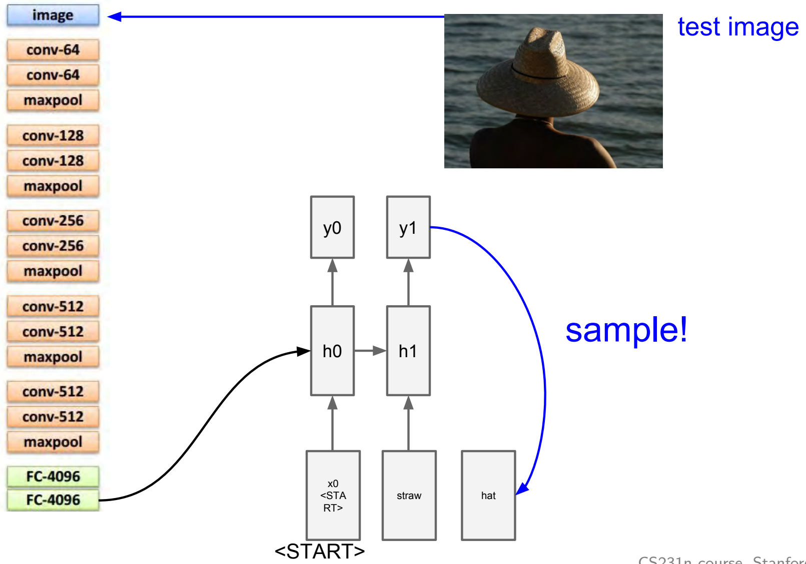
Image Captioning



test image

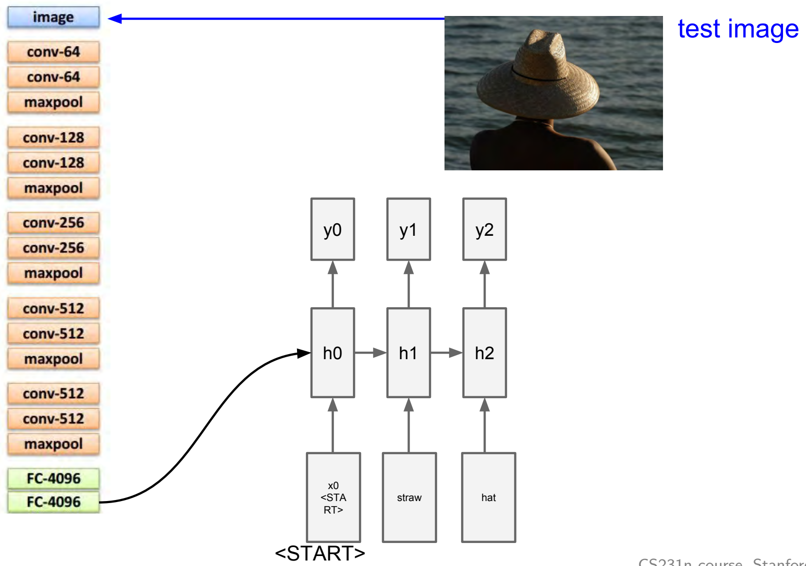
CS231n course, Stanford University

Image Captioning



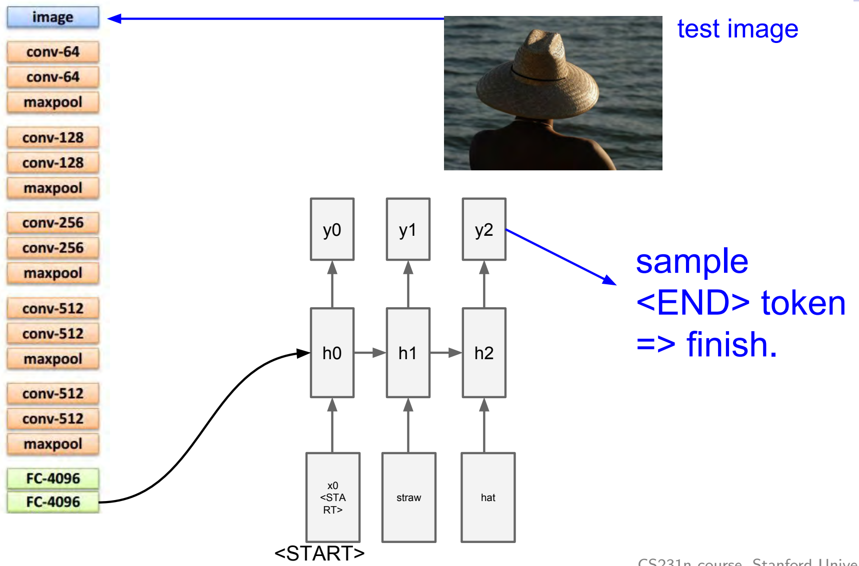
CS231n course, Stanford University

Image Captioning



CS231n course, Stanford University

Image Captioning



CS231n course, Stanford University

Image Captioning

Image Captioning: Example Results

Captions generated using neuraltalk2
All images are CC0 Public domain: cat suitcase, cat tree, dog, bear, surfers, tennis, giraffe, motorcycle

A cat sitting on a suitcase on the floor

A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

A tennis player in action on the court

Two giraffes standing in a grassy field

A man riding a dirt bike on a dirt track

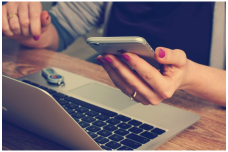
CS231n course, Stanford University

Image Captioning

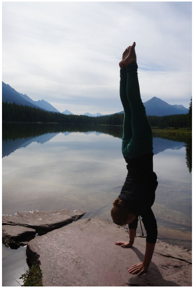
Image Captioning: Failure Cases

Captions generated using neuraltalk2
All images are CC0 Public domain: fur coat, handstand, spider web, baseball

A woman is holding a cat in her hand



A person holding a computer mouse on a desk



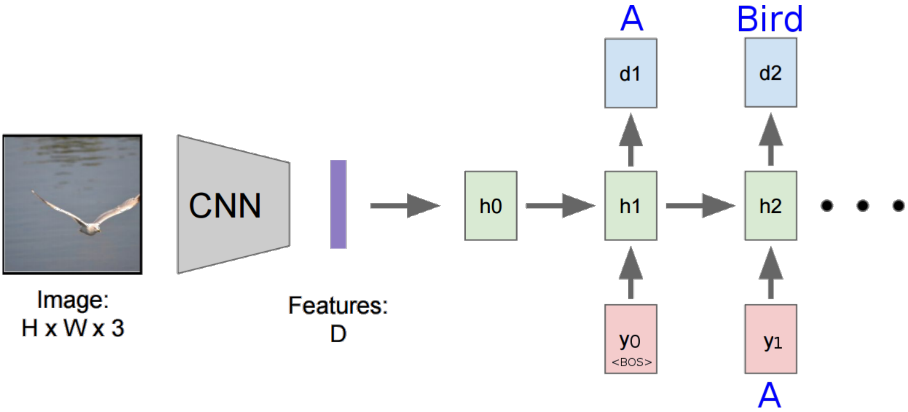
A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

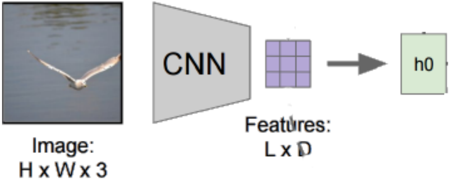
CS231n course, Stanford University

Vanilla Image Captioning



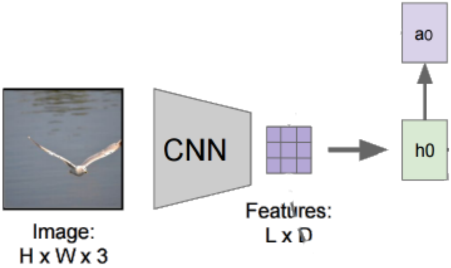
CS231n course, Stanford University

Image Captioning with Attention



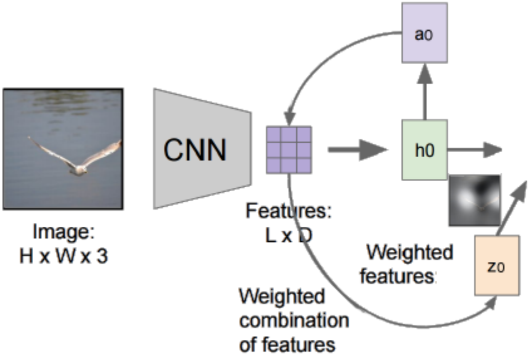
CS231n course, Stanford University

Image Captioning with Attention



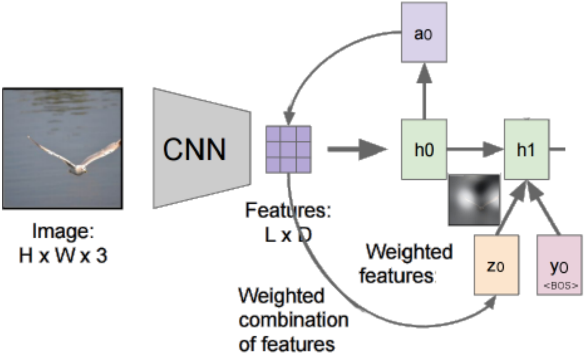
CS231n course, Stanford University

Image Captioning with Attention



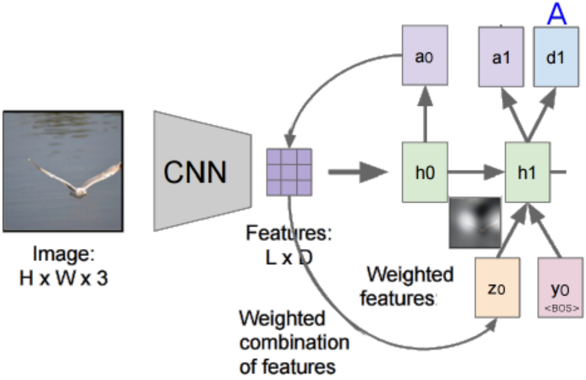
CS231n course, Stanford University

Image Captioning with Attention



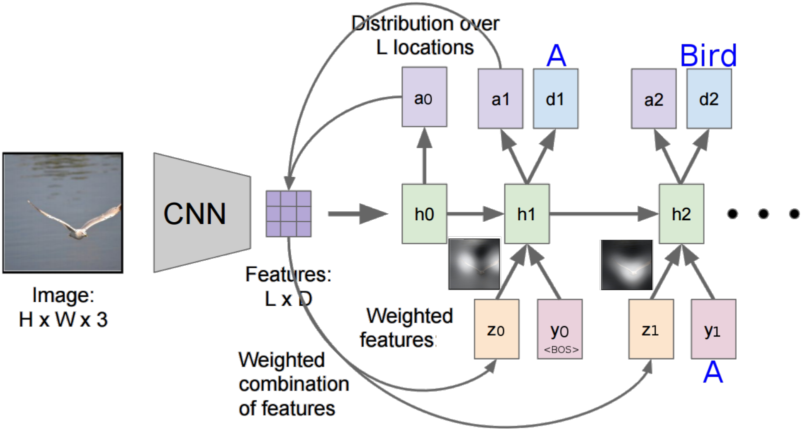
CS231n course, Stanford University

Image Captioning with Attention



CS231n course, Stanford University

Image Captioning with Attention



CS231n course, Stanford University

Video Captioning

§ Example from MSR-VTT Dataset

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.

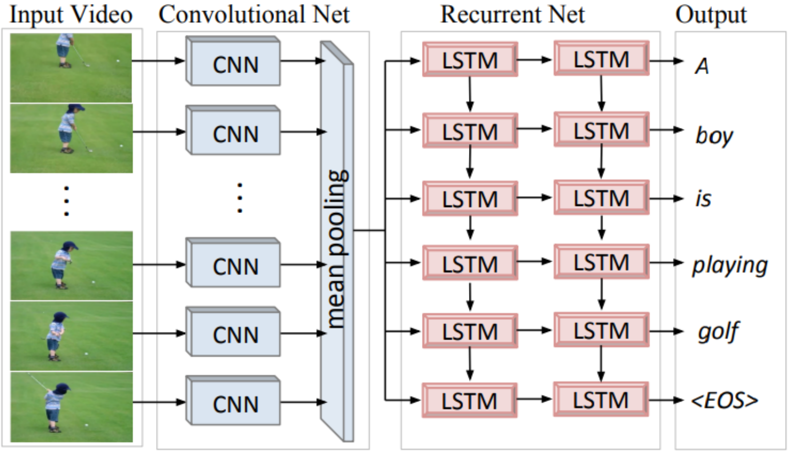
1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

Figure 1. Examples of the clips and labeled sentences in our MSR-VTT dataset. We give six samples, with each containing four frames to represent the video clip and five human-labeled sentences.

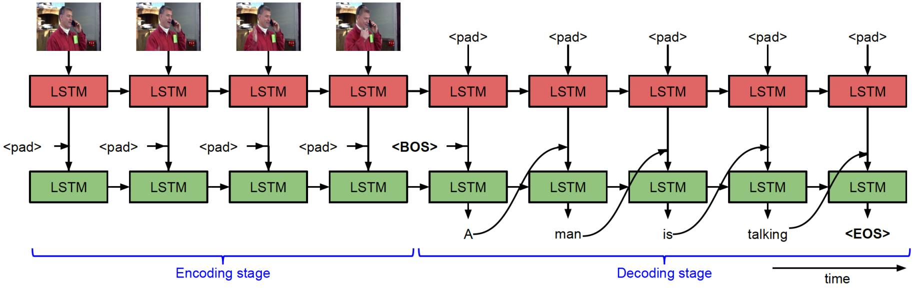
J Xu, T Mei, T Yao and Y Rui, 'MSR-VTT: A Large Video Description Dataset for Bridging Video and Language', CVPR 2016

Video Captioning



S Venugopalan *et al.* 'Translating Videos to Natural Language Using Deep Recurrent Neural Networks', NAACL 2015

Video Captioning



S Venugopalan *et al.* 'Sequence to Sequence – Video to Text', ICCV 2015

Whats Wrong with Deep Captioning Models?

- Deep models doing amazing job on captioning natural images.

- A brown bear standing on top of a lush green field.
  - Donahue *et. al.* - Long-term Recurrent Convolutional Networks for Visual Recognition and Description - CVPR 2015
- A large brown bear walking through a forest.
  - MSR CaptionBot

- Both LRCN and MSR CaptionBot did a pretty good job in describing the image. They are able to get the main object in the image, what it is doing, where it is *etc.*

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Whats Wrong with Deep Captioning Models?

A brown bear walking across a lush green field.

A large brown bear walking through a forest.

A brown bear walks in the grass in front of trees.

A brown bear sitting on top of a green field.

A brown bear walking on a grassy field next to trees.

A large brown bear walking across a lush green field.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Whats Wrong with Deep Captioning Models?

§ Cannot generalize to new objects

§ A black bear is standing in the grass.

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

§ A bear that is eating some grass.

- MSR CaptionBot

§ Deep Compositional Captioner model gives - A anteater is standing in the grass.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Deep Compositional Captioning

Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data

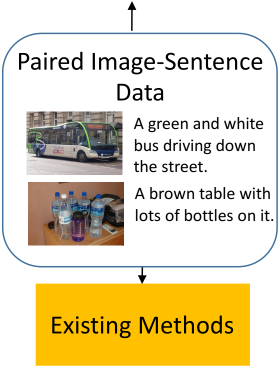
CVPR 2016 (Oral)





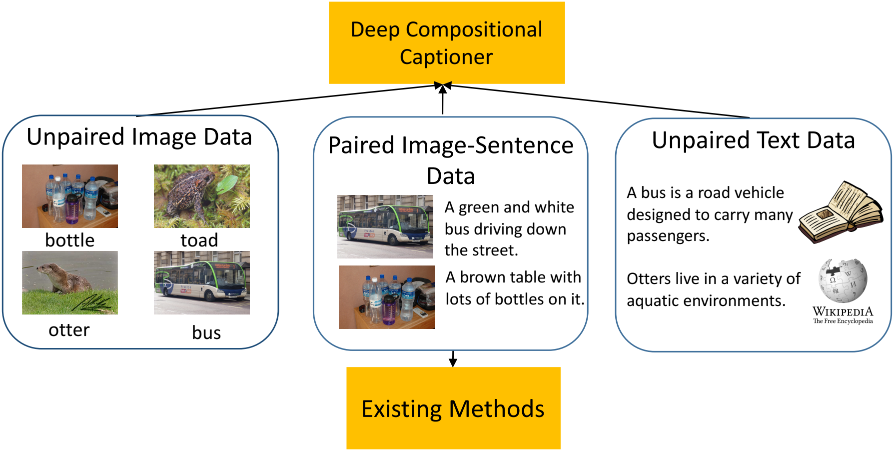

Lisa Hendricks, UC Berkeley
Subhashini Venugopalan, UT Austin
Marcus Rohrbach, UC Berkeley
Raymond Mooney, UT Austin
Kate Saenko, Boston University
Trevor Darrell, UC Berkeley

Deep Compositional Captioning



Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Deep Compositional Captioning



Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Deep Compositional Captioning

- The approach consists of 3 stages.

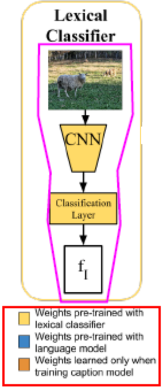


Figure 2: The DCC consists of a lexical classifier, which maps pixels to semantic concepts and is trained only on unpaired image data, and a language model, which learns the structure of natural language and is trained on unpaired text data. The multimodal unit of DCC integrates the lexical classifier and language model and is trained on paired image-sentence data.

- Training a lexical classifier which is nothing but an image classifier.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Deep Compositional Captioning

- The approach consists of 3 stages.

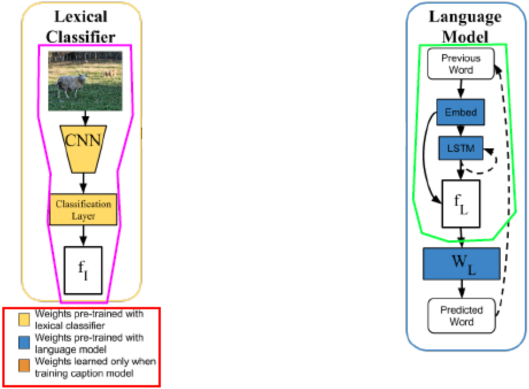


Figure 2: The DCC consists of a lexical classifier, which maps pixels to semantic concepts and is trained only on unpaired image data, and a language model, which learns the structure of natural language and is trained on unpaired text data. The multimodal unit of DCC integrates the lexical classifier and language model and is trained on paired image-sentence data.

- Training a lexical classifier which is nothing but an image classifier.
- Training a language model which predicts a word given previous words in a sentence.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Deep Compositional Captioning

- The approach consists of 3 stages.

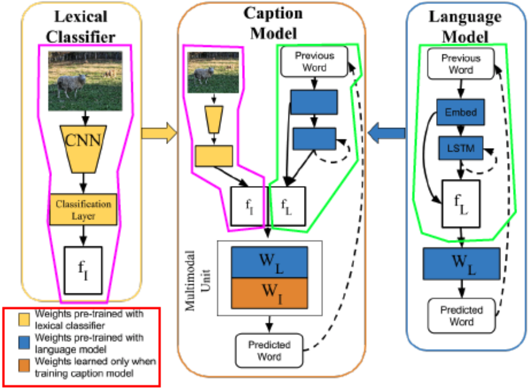
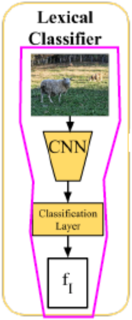


Figure 2: The DCC consists of a lexical classifier, which maps pixels to semantic concepts and is trained only on unpaired image data, and a language model, which learns the structure of natural language and is trained on unpaired text data. The multimodal unit of DCC integrates the lexical classifier and language model and is trained on paired image-sentence data.

- Training a lexical classifier which is nothing but an image classifier.
- Training a language model which predicts a word given previous words in a sentence.
- Combining the lexical classifier and the language model into a captioning model.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Lexical Classifier

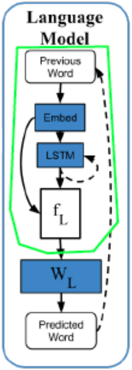


- The lexical classifier is a finetuned CNN on Imagenet dataset.
- The idea behind lexical classifier is more to get visual features for differennt words than to classify objects.

> For example, the sentence "An alpaca stands in the green grass." includes the visual concepts "alpaca", "stands", "green", and "grass". In order to apply multiple labels to each image, we use a sigmoid cross-entropy loss.

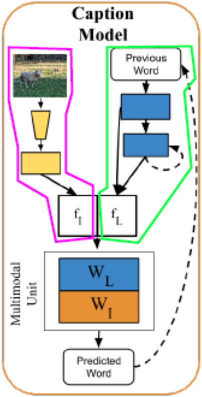
- The output of the lexical classifier is denoted as f_I where each component of f_I corresponds to the probability that a particular concept is present in the image.
- Note that unlike, standard practice where CNN features are taken from an inner layer, it takes output from the output layer.

Language Model



- The language model learns sentence structure using only unpaired data.
- It includes an embedding layer mapping a one-hot vector word representation to a lower dimension space, an LSTM and a word prediction layer.
- It predicts the next word given the previous word.
- The embedded word and the LSTM output are concatenated to form the language features f_L.
- f_L goes through a fully connected layer to output the next word.

Caption Model



- The caption model integrates lexical classifier and language model for image description.
- A multimodal unit combines image features f_I and language features f_L as, $p_w = \text{softmax}(f_I W_I + f_L W_L + b)$, where W_I, W_L and b are learnable parameters.
- Intuitively W_I learn to predict likely words given visual elements discerned by the lexical classifier. W_L learn to predict next word given the previous.
- By summing $f_I W_I$ and $f_L W_L$, the multimodal unit combines the visual information from the lexical classifier and the language knowledge from the language model to form a coherent description.
- Note, W_L is also learned when language model is trained. However, W_I is learned only when image-sentence paired data is available.

Transferring Information Between Objects

Sheep

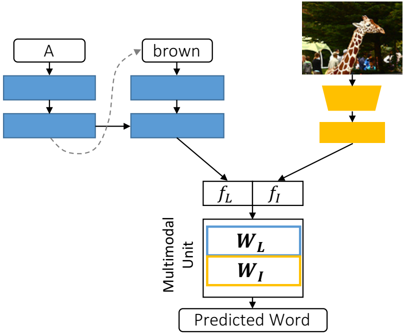
Alpaca

Cake

Scone

Lets Look at an Example

$$p(w_t|I, w_{0:t-1}) = f_L W_L + f_I W_I + b$$



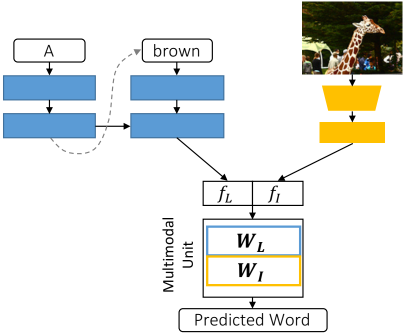
- Say, we want to describe this Giraffe and so far 'A' and 'brown' has been generated.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Lets Look at an Example

$$p(w_t|I, w_{0:t-1}) = f_L W_L + f_I W_I + b$$

$f_L W_L$ large for
Giraffe
Horse
Couch
...
Standing



- Say, we want to describe this Giraffe and so far 'A' and 'brown' has been generated.
- The language features provides these high probability words.

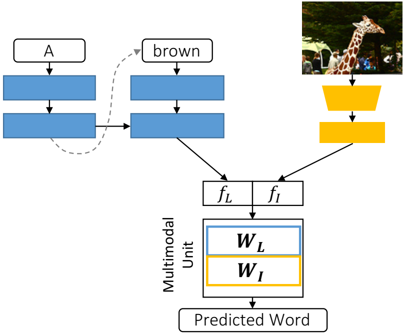
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Lets Look at an Example

$$p(w_t|I, w_{0:t-1}) = f_L W_L + f_I W_I + b$$

$f_L W_L$ large for
Giraffe
Horse
Couch
...
Standing

$f_I W_I$ large for
Giraffe
Trees
Standing
...
Couch



- Say, we want to describe this Giraffe and so far 'A' and 'brown' has been generated.
- The language features provides these high probability words.
- The image features give high probability to words that make sense given the image

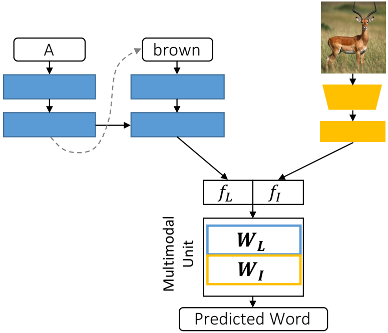
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Lets Look at an Example

$$p(w_t|I, w_{0:t-1}) = f_L W_L + f_I W_I + b$$

$f_L W_L$ large for
Giraffe
Horse
Couch
...
Standing

$f_I W_I$ large for
Giraffe
Trees
Standing
...
Couch



- Now what happens when the 'impala' image is tried to be described.
- This is not in the paired image-sentence dataset.

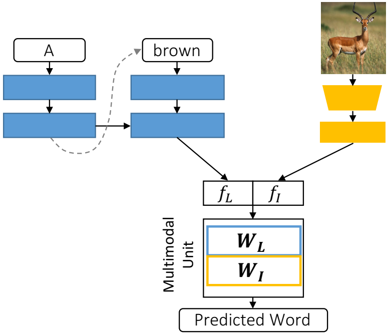
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Lets Look at an Example

$$p(w_t|I, w_{0:t-1}) = f_L W_L + f_I W_I + b$$

$f_L W_L$ large for
Giraffe
Horse
Couch
...
Standing

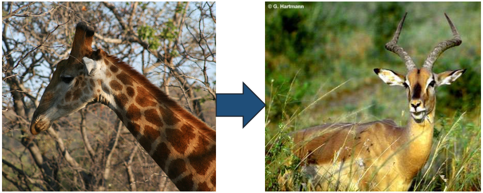
$f_I W_I$ large for
Giraffe
Trees
Standing
...
Couch



- Now what happens when the ‘impala’ image is tried to be described.
- This is not in the paired image-sentence dataset.
- Since, multimodal unit was trained only on the paired image-sentence data, ‘impala’ is still not described.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

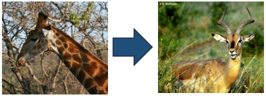


- The authors introduced a transfer mechanism.
- First, used "word2vec" to find a word in the paired image-sentence data and which also similar to "impala".
- In "word2vec" space, it is found that "giraffe" is a word which is most similar to "impala".

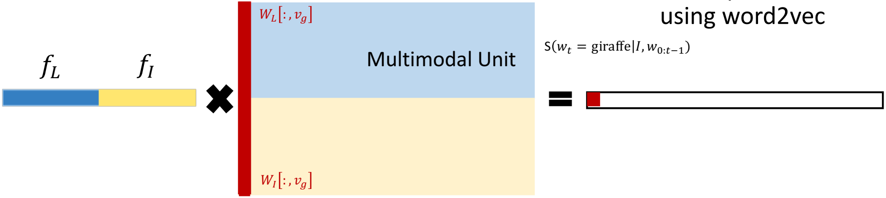
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

$$S(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$



Transfer pair chosen using word2vec



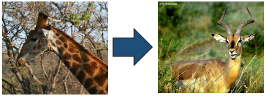
§ The score for giraffe is a linear combination of features and a single column in the multimodal weight matrix.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

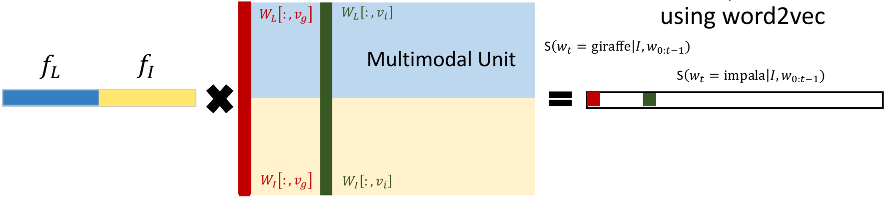
Weight Transfer

$$S(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$

$$S(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec



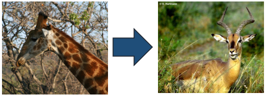
- The score for giraffe is a linear combination of features and a single column in the multimodal weight matrix.
- Similarly the score for impala is a linear combination of features and another single column in the multimodal weight matrix.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

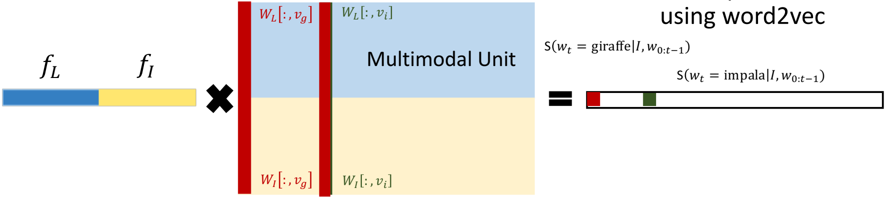
Weight Transfer

$$\mathrm{S}(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$

$$\mathrm{S}(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec

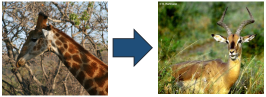


- The score for giraffe is a linear combination of features and a single column in the multimodal weight matrix.
- Similarly the score for impala is a linear combination of features and another single column in the multimodal weight matrix.
- To describe impala similar to girrafe a copy of weights can be made.

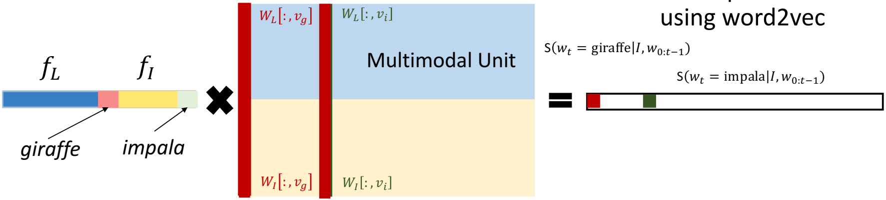
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

$$\mathrm{S}(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$
$$\mathrm{S}(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec

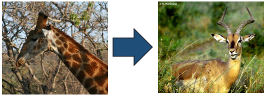


- Remember, the image features here are basically the output layer from the lexical classifier.
- So, we can expect high value for giraffe if there is a giraffe in the image and high value if there is an impala in the image.
- And if we have high image feature value for giraffe we would like to output giraffe as one output word. Similarly for impala too.

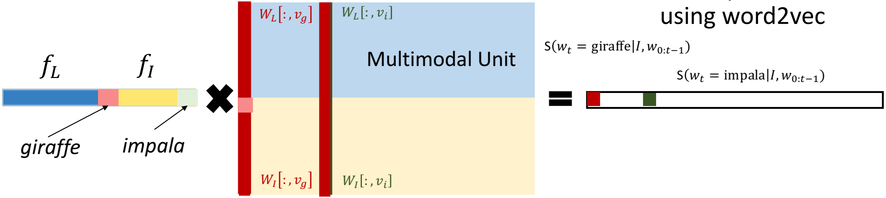
Slide courtesy: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

$$S(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$
$$S(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec

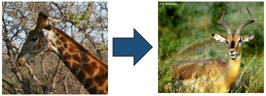


- Now, think about the particular weight that gets multiplied with giraffe feature.
- We would like the impala weight to behave similarly when impala feature is high.

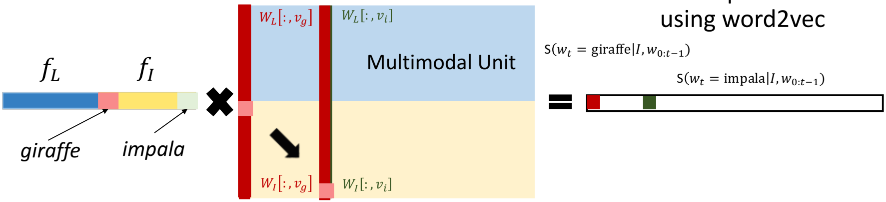
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

$$S(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$
$$S(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec

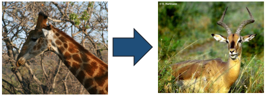


- Now, think about the particular weight that gets multiplied with giraffe feature.
- We would like the impala weight to behave similarly when impala feature is high.
- Thus the giraffe to impala weight transfer also is done.

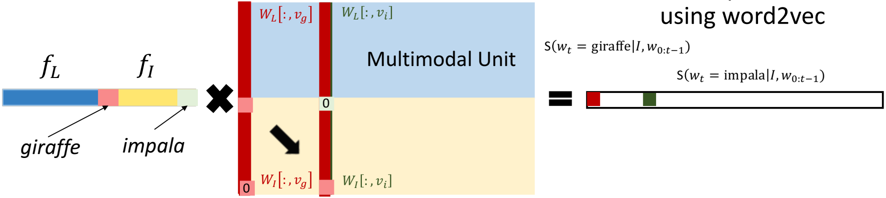
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Weight Transfer

$$\mathrm{S}(w_t = \text{giraffe}|I, w_{0:t-1}) = f_L W_L[:, v_g] + f_I W_I[:, v_g] + b_g$$
$$\mathrm{S}(w_t = \text{impala}|I, w_{0:t-1}) = f_L W_L[:, v_i] + f_I W_I[:, v_i] + b_i$$



Transfer pair chosen using word2vec



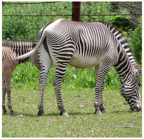
- Now, if we have a giraffe in the image, it should not influence the probability of outputing the impala word and vice-versa.
- So, the corresponding weights are zerod out.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Results

No transfer: A green and white street sign on a city street.
DCC: A green and white **bus** parked on the side of the street.

No transfer: A dog lying on a bed with a large brown dog.
DCC: A dog lying on a **couch** with a large window in the background.



No transfer: Two giraffes are eating grass in the field.
DCC: Two **zebra** grazing in a green grass field.

No transfer: A white and black cat is sitting on a toilet.
DCC: A white **microwave** sitting on a brick wall.

Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Results

DCC can describe over 300 ImageNet visual concepts in diverse contexts.

DCC: A person is holding a **gecko** in their hand.

Berkeley LRCN: A person holding a piece of food in their hand.

MSR CaptionBot: A close up of a person holding a baby.

DCC: A **gecko** is standing on a branch of a tree.

Berkeley LRCN: A bird is standing on the edge of a rock.

MSR CaptionBot: A bird that is standing in the water.

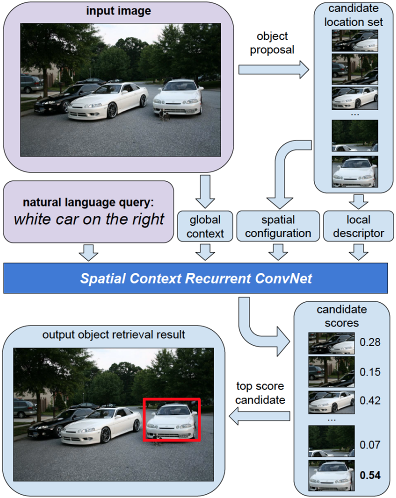
Slide courtesey: Lisa Anne Hendricks, CVPR 2016

Natural Language Object Retrieval

query='*man in middle with blue shirt and blue shorts*'

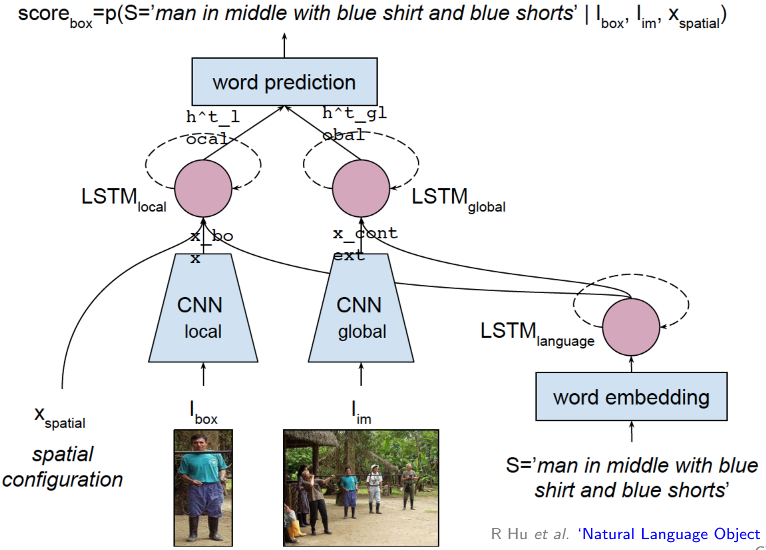
R Hu *et al.* 'Natural Language Object Retrieval', CVPR 2016

Natural Language Object Retrieval



R Hu *et al.* 'Natural Language Object Retrieval', CVPR 2016

Natural Language Object Retrieval



R Hu *et al.* 'Natural Language Object Retrieval', CVPR 2016

Natural Language Object Retrieval

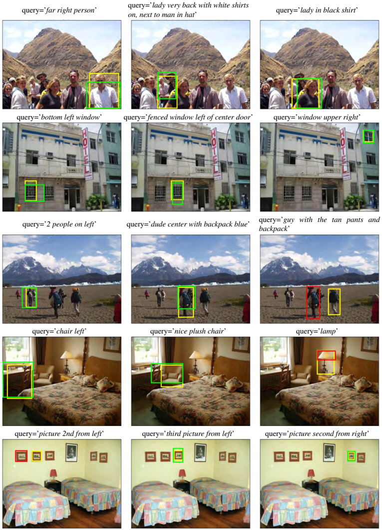
At test time, given an input image I, a query text S and a set of candidate bounding boxes $\{b_i\}$, the query text S is scored on i-th candidate box using the likelihood of the query text sequence conditioned on the local image region, the whole image and the spatial configuration of the box, computed as

$$
\begin{aligned}
s &= p(S|I_{box}, I_{im}, x_{spatial}) &(8)\\
&= \prod_{w_t \in S} p(w_t|w_{t-1}, \cdots, w_1, I_{box}, I_{im}, x_{spatial}) &(9)
\end{aligned}
$$

and the highest scoring candidate boxes are retrieved.

R Hu *et al.* 'Natural Language Object Retrieval', CVPR 2016

Natural Language Object Retrieval



R Hu *et al.* 'Natural Language Object Retrieval', CVPR 2016