Vision and Language
CS60010: Deep Learning

Abir Das

IIT Kharagpur

March 19 and 26, 2020
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

- One to one
- One to many
- Many to one
- Many to many
- Many to many
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
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.

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.
Applications

Image and video retrieval by content

- mountain with trees
Applications

Image and video retrieval by content

Video description service

Children are wearing green shirts. They are dancing as they sing the carol.
Applications

Image and video retrieval by content

- "mountain with trees"

Video description service

- Children are wearing green shirts. They are dancing as they sing the carol.

Human Robot Interaction
Applications

Image and video retrieval by content

Video description service

Children are wearing green shirts. They are dancing as they sing the carol.

Human Robot Interaction

Video surveillance
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.
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

Recurrent Neural Network

Convolutional Neural Network
Image Captioning
Introduction
Captioning
Object Retrieval

Image Captioning

<START>
<START>
test image

Abir Das (IIT Kharagpur)

CS60010
March 19 and 26, 2020 10 / 53
Image Captioning

Before (No Image):
\[ h = \tanh(W_{xh} \cdot x + W_{hh} \cdot h) \]

Now (With Image):
\[ h = \tanh(W_{xh} \cdot x + W_{hh} \cdot h + W_{ih} \cdot v) \]
Image Captioning

<START>

- image
- conv-64
- conv-64
- maxpool
- conv-128
- conv-128
- maxpool
- conv-256
- conv-256
- maxpool
- conv-512
- conv-512
- maxpool
- conv-512
- conv-512
- maxpool
- FC-4096
- FC-4096

y0

h0

x0

<START>

straw

<START>

test image

sample!
Image Captioning
Image Captioning

Abir Das (IIT Kharagpur)

<START>
test image

h0
x0
<START>

y0

y1

h0

h1

x0<br>
<START>

straw

hat

sample!
Image Captioning

- **<START>**
- **test image**

Diagram showing:
- **image**
  - conv-64
  - conv-64
  - maxpool
  - conv-128
  - conv-128
  - maxpool
  - conv-256
  - conv-256
  - maxpool
  - conv-512
  - conv-512
  - maxpool
  - conv-512
  - conv-512
  - maxpool
  - FC-4096
  - FC-4096

- **x0**: <START>
- **y0**: straw
- **y1**: hat
- **y2**: 

**Abir Das (IIT Kharagpur)**

CS231n course, Stanford University

March 19 and 26, 2020 15 / 53
Image Captioning

<START>

test image

sample <END> token => finish.

<START>

x0
<START>
straw
hat

y0
y1
y2

h0
h1
h2

FC-4096
FC-4096

Abir Das (IIT Kharagpur)

CS231n course, Stanford University

March 19 and 26, 2020 16 / 53
Image Captioning: Example Results

A cat sitting on a suitcase on the floor
A cat 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
Image Captioning: Failure Cases

- A woman is holding a cat in her hand
- A woman standing on a beach holding a surfboard
- A person holding a computer mouse on a desk
- A bird is perched on a tree branch
- A man in a baseball uniform throwing a ball
Vanilla Image Captioning

Image: H x W x 3

Features: D

CNN

h0 → h1 → h2

A

d1

Bird

d2

y0
<BOS>
y1

A
Image Captioning with Attention

Image:
H x W x 3

Features:
L x D

CNN

h0
Image Captioning with Attention

![Diagram of Image Captioning with Attention]

- **Image**: \(H \times W \times 3\)
- **CNN**
- **Features**: \(L \times D\)
- **ao**
- **ho**
Image Captioning with Attention
Image Captioning with Attention

Image: $H \times W \times 3$

Features: $L \times D$

Weighted features:

Weighted combination of features

CNN
Image Captioning with Attention

Diagram:
- **Image**: $H \times W \times 3$
- **Features**: $L \times D$
- **Weighted features**
- **Weighted combination of features**

Diagram components:
- **CNN**
- **A**
- **a0**, **a1**, **d1**
- **h0**, **h1**
- **z0**, **y0**
- **<BOS>**
Image Captioning with Attention

- **Introduction**
- **Captioning**
- **Object Retrieval**

**Image:**
- **Features:** \( H \times W \times D \)
- **Weighted combination of features**

**Network Diagram:**
- **CNN**
- **Distribution over \( L \) locations**
- **Weighted features**
- **A**
- **Bird**
- **\( a_0, a_1, d_1, a_2, d_2 \)**
- **\( h_0, h_1, h_2 \)**
- **\( z_0, y_0, z_1, y_1 \)**
- **\( <BOS> \)**

**CS231n course, Stanford University**
Video Captioning

§ Example from MSR-VTT Dataset

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.
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
What's Wrong with Deep Captioning Models?

1. 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

2. 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.
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.
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.
Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data

CVPR 2016 (Oral)
Deep Compositional Captioning

Paired Image-Sentence Data
- A green and white bus driving down the street.
- A brown table with lots of bottles on it.

Existing Methods

Slide courtesy: Lisa Anne Hendricks, CVPR 2016
Deep Compositional Captioning

Existing Methods

Unpaired Image Data
- bottle
- toad
- otter
- bus

Paired Image-Sentence Data
- A green and white bus driving down the street.
- A brown table with lots of bottles on it.

Unpaired Text Data
- A bus is a road vehicle designed to carry many passengers.
- Otters live in a variety of aquatic environments.

Deep Compositional Captioner

Slide courtesy: Lisa Anne Hendricks, CVPR 2016
Deep Compositional Captioning

The approach consists of 3 stages.

- 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.

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.
Deep Compositional Captioning

The approach consists of 3 stages.

- 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.

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.
The approach consists of 3 stages.

- 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.

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.
Lexical Classifier

- The lexical classifier is a finetuned CNN on Imagenet dataset.
- The idea behind lexical classifier is more to get visual features for different 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.
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.
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
\[ p(w_t | l, w_{0:t-1}) = f_L W_L + f_I W_I + b \]

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

Slide courtesy: Lisa Anne Hendricks, CVPR 2016
**Let’s Look at an Example**

\[ p(w_t | l, w_{0:t-1}) = f_L W_L + f_I W_I + b \]

- \( f_L W_L \) 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 provide these high probability words.

*Slide courtesy: 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...
- $f_I W_I$ large for Giraffe, Trees, 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.

The image features give high probability to words that make sense given the image.

Slide courtesy: Lisa Anne Hendricks, CVPR 2016
Let's Look at an Example

\[ p(w_t|l,w_{0:t-1}) = f_LW_L + f_IW_I + b \]

\( f_LW_L \) large for
- Giraffe
- Horse
- Couch

... Standing

\( f_IW_I \) 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.
Let's 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.
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”.
Weight Transfer

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

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

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

- 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 giraffe a copy of weights can be made.

Transfer pair chosen using word2vec

Slide courtesy: Lisa Anne Hendricks, CVPR 2016
Weight Transfer

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

- 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.

Transfer pair chosen using word2vec

\( S(w_t = \text{giraffe}|I, w_{0:t-1}) \)
\( S(w_t = \text{impala}|I, w_{0:t-1}) \)
Weight Transfer

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

§ 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.

Transfer pair chosen using word2vec

\[ S(w_t = \text{giraffe}|I, w_{0:t-1}) \]
\[ S(w_t = \text{impala}|I, w_{0:t-1}) \]
Weight Transfer

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

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.
Weight Transfer

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

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 zeroed out.
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 courtesy: Lisa Anne Hendricks, CVPR 2016
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.
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
score_{box} = p(S='man in middle with blue shirt and blue shorts' | I_{box}, I_{im}, X_{spatial})

R Hu et al. ‘Natural Language Object Retrieval’, CVPR 2016
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

$$ s = p(S|I_{box}, I_{im}, x_{spatial}) $$ (8)

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

and the highest scoring candidate boxes are retrieved.

R Hu et al. ‘Natural Language Object Retrieval’, CVPR 2016
Natural Language Object Retrieval

Figure 7. Examples on multiple objects in the same image in ReferIt, showing the highest scoring candidate box (correct in green, incorrect in red) from 100 EdgeBox proposals and ground truth (yellow). Our model retrieves the objects by taking their local descriptors, spatial configurations and scene-level contextual information into account.

R Hu et al. ‘Natural Language Object Retrieval’, CVPR 2016