IMAGE CAPTIONING USING PHRASE-BASED HIERARCHICAL LSTM MODEL

Chee Seng Chan PhD SMIEEE
23 October 2017
Nvidia AI Conference, Singapore
email: cs.chan@um.edu.my
Aim: Automatic generate a full sentence describing an image.

Motivated by the significant progress of image classification and statistical language model.

Applications:
- Early childhood educations
- Scene understanding for the visual impairments
- Image retrievals

Two children are playing on a swing made out of a tire.
BACKGROUNDs

- Processing of Image, $I$:
  - Represented as a vector using feature learning algorithm such as convolutional neural network (CNN)

- Processing of Language:
  - Each sentence is equivalent to a sequence of words.
  - A statistical model is trained to predict the conditional probability of next word given all previous words

$$P(w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1})$$

- Multimodal Embedding
  - Prediction of next word also conditioned on image

$$P(w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, I)$$
Sequence is learned with Recurrent Neural Network (RNN).

The most popular variant of RNN is Long Short-Term Memory (LSTM).
Problem Statement

- Conventional models treat a sentence as a sequence of words.
- All other linguistic syntax and structure are disregarded.
- Sentence structure is one of the most prominent characteristic of sentence!

Two dogs are running in the snow.

NP VP PP NP

NP = noun phrase
VP = verb phrase
PP = prepositional phrase
PROBLEM STATEMENT

- Quoted on Victor Yngve \(^{[14]}\) (an influential contributor in linguistic theory):
  
  "language structure involving, in some form or other, a phrase structure hierarchy, or immediate constituent organization"

- Example:
Is it really okay to treat a sentence as only a sequence of words, while disregarding any other important characteristic of sentence such as structure?

1. Design of phrase-based model for image captioning. This is one of the most earliest work after PbIC[13].
2. Investigate on its performance as compared to a pure sequence model.
Design motivation

- Noun phrases form most of an image caption.
- They have similar syntactic role
- They have strong relation with the image.

A young girl wearing a yellow shirt with a blue backpack is walking next to a fence covered with a blue plastic cover.
Sentence: A motorcyclist on the street.
# Related Works

<table>
<thead>
<tr>
<th>Methods</th>
<th>Details (Red words are their cons)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template based</td>
<td>• Generate sentence from a fixed template.</td>
<td>1-4</td>
</tr>
<tr>
<td></td>
<td>• Sentence generated is rigid.</td>
<td></td>
</tr>
<tr>
<td>Composition Method</td>
<td>• Stitch up image relevant phrases to form a sentence.</td>
<td>5-7</td>
</tr>
<tr>
<td></td>
<td>• Computational cost is high.</td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>• Trained to predict sequence.</td>
<td>mRNN [8],</td>
</tr>
<tr>
<td></td>
<td>• Only model words sequence.</td>
<td>NIC [9],</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DeepVS [10],</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LCRNN [12]</td>
</tr>
</tbody>
</table>

- The closest work is “Phrase based Image Captioning” PbIC[13] proposed by Lebret et al.
- They encode each sentence as phrase sequence only while my proposal is to encode as sequence of phrase and words.
- They use simpler model.
PROPOSED MODEL

- Training Data: image sentence pair
**Proposed Model:**

1) **Phrase Chunking**

- Approach to identify the constituents of a sentence.
- Extract only *noun phrase* – prominent in image description
- **Dependency parse*** with selected relations:
  - **det** – determiner (e.g.: “a man”)
  - **amod** - adjective modifier (e.g.: “green shirt”)
  - **nummod** - numeric modifier (e.g.: “two dogs”)
  - **compound** - compound (e.g.: “basketball court”)
  - **advmod** - adverbial modifier, when modifying meaning of adjective (e.g.: “dimly lit room”)
  - **nmod:of & nmod:poss** - nominal modifier for possessive alteration (e.g.: “his hand”)

*Stanford CoreNLP Software - https://stanfordnlp.github.io/CoreNLP/*
PROPOSED MODEL: 
1) PHRASE CHUNKING

- Chunking from dependency parse

```
The man in the gray shirt and sandals is pulling the large tricycle
```

```
selective dependency parsing
```

```
det(man, the)  amod(shirt, gray)  amod(tricycle, large)
det(shirt, the)  det(tricycle, the)
```

```
Our refinement (proposed)
```

```
The man in the gray shirt and sandals is pulling the large tricycle
```

```
New Chunking
```

```
the man the gray shirt sandals the large tricycle
```
PROPOSED MODEL:
2) COMPOSITIONAL VECTOR OF PHRASE

- Our proposed architecture is the hierarchical counterpart of NIC model proposed by Vinyals et al [9]

![Diagram showing the hierarchical structure of the proposed model with LSTM layers and compositional phrase vector.]
PROPOSED MODEL: 3) SENTENCE ENCODING

Sentence:
The man in the gray shirt and sandals is pulling the large tricycle.

- A ‘phrase’ token is added into the corpus for prediction
Objective function:

\[
C_F(\theta) = -\frac{1}{L} \sum_{j=1}^{M} [N_j \log_2 P_{\mathcal{P}\mathcal{L}(S_j|I_j)} + C_{PS_j}] + \lambda_0 \| \theta \|^2_2 \quad L = M \times \sum_{j=1}^{M} N_j .
\]

\( j / M = \text{index / total no of training sentence) \)

Perplexity:

\[
\log_2 P_{\mathcal{P}\mathcal{L}(S|I)} = -\frac{1}{S} \sum_{t_a=-1}^{S} \log_2 p_{t_s}
\]

\[
\log_2 P_{\mathcal{P}\mathcal{L}(S|I)} = -\frac{1}{N} \left[ \sum_{t_a=-1}^{Q} \log_2 p_{t_s} + \sum_{i=1}^{R} \left[ \sum_{t_p=-1}^{P_i} \log_2 p_{t_p} \right] \right] , \quad N = Q + \sum_{i=1}^{R} P_i .
\]

\( p_{t_p} / p_{t_s} = \text{probability distribution over words on the particular time step for phrase / sentence} \)

\( t_p / P = \text{time step / total no. of time step in phrase} \)

\( t_s / Q = \text{time step / total no. of time step in sentence} \)

\( i / R = \text{index / total no. of phrase in sentence I} \)
**TRAINING — PHRASE SELECTION OBJECTIVE**

- Objective function:

\[
C_F(\theta) = -\frac{1}{L} \sum_{j=1}^{M} [N_j \log_2 \mathcal{P}_j \mathcal{L}(S_j|I_j) + C_{PS,j}] + \lambda_\theta \cdot \| \theta \|_2^2
\]

- Cost of phrase selection objective:

\[
C_{PS} = \sum_{t_a \in \mathcal{P}} \sum_{k=1}^{H} \kappa_{t_a,k} \sigma(1 - y_{t_a,k} h_{t_a,k} W_{\text{ps}})
\]

- \(W_{\text{ps}}\) = trainable parameters
- \(h_{t_a,k}\) = hidden output at \(t_a\) for input \(k\)
- \(y_{t_a,k}\) = label of input \(k\) at \(t_a\)
- \(\kappa_{t_a,k}\) = normalizing constant based on
- \(k / H\) = index / total no of inputs at \(t_a\)
- \(\mathcal{P}\) = set of \(t_a\) which the input is phrase
GRAPHICAL ILLUSTRATION: SENTENCE GENERATION (PHRASE LEVEL)

Selected Phrases:
- two dogs
- the snow
- the beach
- a dog

Selected Phrases:
- a brown dog
- two dogs
- the snow
- the beach
- a dog

CNN & Img Embedding

LSTM

#START#

K candidates

K^2 candidates

a
the
two
its
three
dogs
snow
brown
beach
dog
dog


two
its
three

a brown dog

K candidates
**GRAPHICAL ILLUSTRATION: SENTENCE GENERATION (SENTENCE LEVEL)**

Two dogs play in the snow.

*K Selected Phrases:*
- a brown dog
- two dogs
- the snow
- the beach
**Experiment**

- Tested on Flickr8k and Flickr30k datasets.
- Each image is annotated with five descriptions by human.
- 1k of images are used for validation and another 1k of images are used for testing, while the rest are for training (consistent with state-of-the-art).

- A woman in a red coat with a man in a white and black coat and a black dog in the snow.
- Two people and a dog are in the snow.
- Two people are interacting with a dog that has bitten an object one of them is holding.
- Two people are walking up a snowy hill with a dog.
- Two people playing on a snowy hill.
QUALITATIVE RESULTS (PHRASE)

- Phrase generation:
QUALITATIVE RESULTS (SENTENCE)

Image:

NIC: (baseline)
A skateboarder does a trick on a ramp.

phi-LSTM (proposed)
A man doing a trick on a bike.

Reference: (human)
A skateboarder on a ramp.

A man on a snowy mountain.

A person in the snow.

A surfer rides a wave.

A person in the water.

A man crouched on a snowy peak.

A surfer does a flip on a wave.

Image:

NIC: (baseline)
A group of people are standing in front of a building.

Three people are standing in front of three men.

Reference: (human)
A group of tourists stand around as a lady puts her hand near the mouth of a statue.

A man is doing a trick on a skateboard.

A skateboarder does a trick on a ramp.

A skateboarder in the air at a big outdoor ramp.

Two dogs play in the grass.

Three dogs play in a grassy field.

The three dogs ran in the yard.
MORE RESULTS (SENTENCES WITH SAME OBJECT(S))

Dog
- Two dogs play in a grassy field.
- A dog in a race.
- A small dog jumps to catch a toy.

Action
- A snowboarder in the air.
- A skateboarder does a trick on a skateboard.
- A person does a trick on a bicycle.

Human
- A person in a helmet is riding a dirt bike.
- A surfer in a wave.
- A young boy jumps into a swimming pool.

Human
- A group of women in the camera.
- A little boy in a car.
- A child in a swing.
MORE RESULTS (SENTENCES WITH SAME SCENE)

Places

A group of people in the snow.
A woman in the snow.
A woman in the street.

A group of people in a field.
A person is riding a dirt bike.
A man is riding a bike.

A girl in the water.
A man in the water.
A surfer in the water.
Qualitative Results (Poor examples)

A man in a boat in the water.
A child in a slide.
A woman is holding a young boy.
A woman and a child are sitting in a baby.
A woman in a man in a kitchen.
A man is holding a woman.
Quantitative Results

- Evaluation metric: BLEU
- Measure n-grams **precision** quality between generated caption and reference sentences (human).

<table>
<thead>
<tr>
<th>Models</th>
<th>Flickr8k</th>
<th>Flickr30k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-1  B-2 B-3 B-4</td>
<td>B-1  B-2 B-3 B-4</td>
</tr>
<tr>
<td>DeepVS [4]</td>
<td>57.9  38.3 24.5 16.0</td>
<td>57.3  36.9 24.0 15.7</td>
</tr>
<tr>
<td>NIC [3]</td>
<td>60.2(63) 40.4 25.9 16.5</td>
<td>66.3(66) 42.3 27.7 18.3</td>
</tr>
<tr>
<td>phi-LSTM</td>
<td><strong>63.6</strong>  <strong>43.6</strong>  <strong>27.6</strong>  <strong>16.6</strong></td>
<td><strong>66.6</strong>  <strong>45.8</strong>  <strong>28.2</strong>  <strong>17.0</strong></td>
</tr>
</tbody>
</table>

Our proposed model
MORE ANALYSIS BY COMPARING WITH BASELINE

- Given same amount of training data, and same set of test image, and same set of setting in training:
  - Our model can generate sentence formed with more variety of words in the training corpus.

- What is the minimum time a word should appear in training data, so the model can generate sentence using that word?
  - Our model (phi-LSTM) = 81
  - Baseline (NIC) = 93
CONCLUSION

- Proposed of hierarchical phrase-based LSTM model to generate image description.

- Hierarchical model vs pure sequential model:
  - Able to generate better description
  - Can learn with less data

- Published in ACCV 2016, extension to journal.

- Future works
  - Experiments on MSCOCO dataset
  - Evaluation on more types of automatic evaluation metrics such as ROUGE, METEOR, CIDEr
  - Apply on image sentence bi-directional retrieval
  - Tackle problem in poor results
REFERENCES


REFERENCES


REFERENCES


THE END

Q & A?

Chee Seng Chan Ph.D SMIEEE
University of Malaya, Malaysia
www.cs-chan.com