Phrase-based Image Captioning

Rémi Lebret, Pedro O. Pinheiro, Ronan Collobert

Idiap Research Institute / EPFL

ICML, 9 July 2015
Objective: Generate descriptive sentences given a sample image.

A man is grinding a ramp on a skateboard.
Related Works

- Recent models based on Deep CNN + RNN [Vinyals et al., Karpathy & Fei-Fei, Mao et al., Donahue et al.].

A man is grinding a ramp on a skateboard.
Related Works

- Recent models based on Deep CNN + RNN [Vinyals et al., Karpathy & Fei-Fei, Mao et al., Donahue et al.].

A man is grinding a ramp on a skateboard.

Visual features with Deep CNN
Related Works

- Recent models based on Deep CNN + RNN [Vinyals et al., Karpathy & Fei-Fei, Mao et al., Donahue et al.].
Related Works

- Recent models based on Deep CNN + RNN [Vinyals et al., Karpathy & Fei-Fei, Mao et al., Donahue et al.].

Can similar performance be achieved with a simpler model?
Syntax Analysis of Image Descriptions

A given image $i \in I$

Ground-truth descriptions $s \in S$:

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skateboard
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick
Syntax Analysis of Image Descriptions

A given image \( i \in I \)

Ground-truth descriptions \( s \in S \):

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skate board
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick

→ Chunking approach to identify the sentence constituents.
Syntax Analysis of Image Descriptions

A given image $i \in \mathcal{I}$

Ground-truth descriptions $s \in \mathcal{S}$:

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skate board
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick

→ Chunking approach to identify the sentence constituents.
Syntax Analysis of Image Descriptions

A given image $i \in I$

Ground-truth descriptions $s \in S$:

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skateboard
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick

→ Chunking approach to identify the sentence constituents.
A given image $i \in I$

Ground-truth descriptions $s \in S$:

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skate board
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick

→ Chunking approach to identify the sentence constituents.
A given image $i \in I$

Ground-truth descriptions $s \in S$:

- a man riding a skateboard up the side of a wooden ramp
- a man is grinding a ramp on a skateboard
- man riding on edge of an oval ramp with a skateboard
- a man in a helmet skateboarding before an audience
- a man on a skateboard is doing a trick

→ Chunking approach to identify the sentence constituents.
Syntax Analysis of Image Descriptions

A given image \( i \in I \)

Ground-truth descriptions \( s \in S \):

- A man riding a skateboard up the side of a wooden ramp
- A man is grinding a ramp on a skateboard
- Man riding on edge of an oval ramp with a skateboard
- A man in a helmet skateboarding before an audience
- A man on a skateboard is doing a trick

Key elements in images:

> Noun phrases (NP)
> Verbal phrases (VP)
> Prepositional phrases (PP)

Interactions between elements.
Large-scale Syntax Analysis

- Two datasets: Flickr30k + COCO (≈ 560k training sentences).

- Describing images:
  1. Predicting NP, VP and PP.
  2. Finding how they all interact.
Our approach:

1. A bilinear model that learns a metric between an image and phrases used to describe it.

2. Sentences generated using a simple language model based on caption syntax statistics.
A Bilinear Model $U^T V$

$I = \text{set of training images}$

$C = \text{set of all phrases used to describe } I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$

$V \in \mathbb{R}^{m \times n}$

$\{ \text{trainable parameters } \theta \}$

A man in a helmet skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.
A Bilinear Model $U^T V$

$I$ = set of training images
$C$ = set of all phrases used to describe $I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$
$V \in \mathbb{R}^{m \times n}$

trainable parameters $\theta$

pre-trained CNN representation $z_i \in \mathbb{R}^n$

A man in a helmet skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.

A man
a skate board
a wooden ramp
riding
is grinding
on
with
A Bilinear Model $U^T V$

$I = \text{set of training images}$
$C = \text{set of all phrases used to describe } I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$
$V \in \mathbb{R}^{m \times n}$ \{trainable parameters $\theta$\}

pre-trained CNN representation $z_i \in \mathbb{R}^n$

representation $u_c$ for a phrase $c = \{w_1, \ldots, w_K\}$ by averaging pre-trained word vector representations $x_w \in \mathbb{R}^m$:

$$u_c = \frac{1}{K} \sum_{k=1}^{K} x_{w_k}$$

A man in a helmet skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.
A Bilinear Model $U^T V$

$I = \text{set of training images}$
$C = \text{set of all phrases used to describe } I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$
$V \in \mathbb{R}^{m \times n}$

\{ trainable parameters $\theta$ \}

pre-trained CNN representation $z_i \in \mathbb{R}^n$

representation $u_c$ for a phrase $c = \{w_1, \ldots, w_K\}$ by averaging pre-trained word vector representations $x_w \in \mathbb{R}^m$:

$$u_c = \frac{1}{K} \sum_{k=1}^{K} x_{w_k}$$

A man in a helmet skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.
A Bilinear Model $U^T V$

$I$ = set of training images
$C$ = set of all phrases used to describe $I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$
$V \in \mathbb{R}^{m \times n}$

$\theta$ trainable parameters

pre-trained CNN representation $z_i \in \mathbb{R}^n$

representation $u_c$ for a phrase $c = \{w_1, \ldots, w_K\}$ by averaging pre-trained word vector representations $x_{w} \in \mathbb{R}^m$:

$$u_c = \frac{1}{K} \sum_{k=1}^{K} x_{w_k}$$
A Bilinear Model $U^T V$

$I = \text{set of training images}$
$C = \text{set of all phrases used to describe } I$

$U = (u_{c_1}, \ldots, u_{c_{|C|}}) \in \mathbb{R}^{m \times |C|}$ \text{ trainable parameters } \theta$

$V \in \mathbb{R}^{m \times n}$

score between the image $i$ and a phrase $c$: $f_\theta(c, i) = u_c^T V z_i$

pre-trained CNN representation $z_i \in \mathbb{R}^n$

representation $u_c$ for a phrase $c = \{w_1, \ldots, w_K\}$ by averaging pre-trained word vector representations $x_w \in \mathbb{R}^m$:

$$u_c = \frac{1}{K} \sum_{k=1}^{K} x_{w_k}$$

A man in a helmet skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.
A Bilinear Model $U^T V$

Training with negative sampling by minimizing this logistic loss function w.r.t. $\theta$:

$$\theta \leftarrow \sum_{i \in I} \sum_{c_j \in C^i} \left( \log \left( 1 + e^{-u_{c_j}^T V z_i} \right) + \sum_{c_k \in C^-} \log \left( 1 + e^{+u_{c_k}^T V z_i} \right) \right)$$

→ Stochastic gradient descent, new set of negative phrases $C^-$ at each iteration.
From Phrases to Sentence

▶ Bilinear model gives the $L$ most likely phrases $c_j$.
▶ Generating sentences from this set using $l \in \{1, \ldots, L\}$ phrases:

$$P(c_1, c_2, \ldots, c_l) = \prod_{j=1}^{l} P(c_j|c_1, \ldots, c_{j-1})$$

$$\approx \prod_{j=1}^{l} P(c_j|c_{j-2}, c_{j-1}) \rightarrow 2^{\text{nd}}\text{-order Marchov Chain}$$

▶ Prior knowledge on chunking tags $t \in \{NP, VP, PP\}$:

$$P(c_1, c_2, \ldots, c_l) = \prod_{j=1}^{l} \sum_t P(c_j|t_j = t, c_{j-2}, c_{j-1})P(t_j = t|c_{j-2}, c_{j-1})$$

$$= \prod_{j=1}^{l} P(c_j|t_j, c_{j-2}, c_{j-1})P(t_j|c_{j-2}, c_{j-1})$$
Sentence Decoding

Constrained language model with $t \in \{NP, VP, PP\}$:

\[
P(c_1, c_2, \ldots, c_l) = \prod_{j=1}^{l} P(c_j|t_j, c_{j-2}, c_{j-1}) P(t_j|c_{j-2}, c_{j-1})
\]

→ Beam search to find all $M$ sentences with top $L$ phrases.
Sentence Re-ranking

- Ranking to find the sentence which is the closest to sample image.

- Leveraging score between the image $i$ and a phrase $c$: $f_\theta(c, i) = u_c^T V z_i$.

- Averaging phrase scores $f_\theta(c_j, i) \forall c_j \in s$:

$$\frac{1}{|s|} \sum_{c_j \in s} f_\theta(c_j, i).$$

→ Best candidate = sentence with the highest score.
Experimental Setup

Image dataset:
- **COCO dataset**: 82783/5000/5000 images, 5 sentences per image.
- Only phrases occurring at least 10 times:
  - 8,982 NP (73%)
  - 3,083 VP (75%)
  - 189 PP (99%)

Bilinear model:
- **Image features**: VGG ConvNet pre-trained on Imagenet (4096D vector).
- **Word features**: Hellinger PCA of a word co-occurrence matrix, built over English Wikipedia (400D vector).
- **Trainable parameters** $\theta$:
  - $V \in \mathbb{R}^{400 \times 4096}$ → initialized randomly.
  - $U \in \mathbb{R}^{400 \times |\mathcal{C}|}$ → initialized by averaging word features + fine-tuned.
- 15 negative samples.

Statistical language model:
- Transition probabilities between phrases from COCO dataset.
- **No smoothing**.
- Subset of top-ranked phrases: 20 best NP, 5 best VP and 5 best PP.
## Full Sentence Generation

<table>
<thead>
<tr>
<th>Captioning Method</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human agreement</td>
<td>0.68</td>
<td>0.45</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td><strong>CNN/RNN based models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mao et al.</td>
<td>0.67</td>
<td>0.49</td>
<td>0.35</td>
<td>0.25</td>
</tr>
<tr>
<td>Karpathy &amp; Fei-Fei</td>
<td>0.63</td>
<td>0.45</td>
<td>0.32</td>
<td>0.23</td>
</tr>
<tr>
<td>Vinyals et al.</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Donahue et al.</td>
<td>0.63</td>
<td>0.44</td>
<td>0.30</td>
<td>0.21</td>
</tr>
<tr>
<td><strong>Our model</strong></td>
<td>0.73</td>
<td>0.50</td>
<td>0.34</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Successful example

a bunch of kites flying in the sky on the beach
Successful example

a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group
a bunch of kites flying in the sky on the beach

**NP:** the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group

**VP:** flying, flies, is flying, flying in, are
Successful example

a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group

VP: flying, flies, is flying, flying in, are

PP: on, of, with, in, at
Successful example

A bunch of kites flying in the sky on the beach

**NP:** the beach, a beach, a kite, kites, the ocean, the water, the sky, **people**, a sandy beach, a group

**VP:** flying, flies, is flying, flying in, are

**PP:** on, of, with, in, at

People
Successful example

a bunch of kites flying in the sky on the beach

**NP:** the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group

**VP:** flying, flies, is flying, flying in, are

**PP:** on, of, with, in, at

People flying
Successful example

a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group
VP: flying, flies, is flying, flying in, are
PP: on, of, with, in, at

People flying kites
Successful example

a bunch of kites flying in the sky on the beach

**NP:** the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group

**VP:** flying, flies, is flying, flying in, are

**PP:** on, of, with, in, at

People flying kites on
a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group
VP: flying, flies, is flying, flying in, are
PP: on, of, with, in, at

People flying kites on the beach
Successful example

a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group
VP: flying, flies, is flying, flying in, are
PP: on, of, with, in, at

People flying kites on the beach
a bunch of kites flying in the sky on the beach

**NP:** the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group

**VP:** flying, flies, is flying, flying in, are

**PP:** on, of, with, in, at

**People flying kites on the beach**
Successful example

a bunch of kites flying in the sky on the beach

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group
VP: flying, flies, is flying, flying in, are
PP: on, of, with, in, at

People flying kites on the beach
a flock of geese are walking in a parking lot
a flock of geese are walking in a parking lot

**NP:** a parking lot, parked cars, a black car, car, the road, a street, people, a group, geese, trees

**VP:** parked on, sitting in, driving down, is parked in, crossing

**PP:** of, on, by, in, next to
a flock of geese are walking in a parking lot

NP: a parking lot, parked cars, a black car, car, the road, a street, people, a group, geese, trees
VP: parked on, sitting in, driving down, is parked in, crossing
PP: of, on, by, in, next to

car sitting in a parking lot with parked cars
### Phrase Representation Fine-Tuning

<table>
<thead>
<tr>
<th>PHRASES</th>
<th>NEAREST NEIGHBORS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>A GREY CAT</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>HOME PLATE</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>A HALF PIPE</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>
Conclusion

- Generate image caption by inferring phrases that best describe them.
- Simple model and very fast to train/test.
- We achieve results similar to CNN+RNN models.
- Enriching phrase representations with visual features.

Future research directions:

- Leveraging unsupervised data
- More complex language models
NP: a sign, sky, your attention, cloud, a plane, a cloudy sky, that, a street sign, cloud, you
VP: thank, thanks, flying, sitting in, thanking for
PP: for, on, with, in, next to

Thank you for your attention