Phrase-based Image Captioning

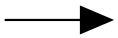
Rémi Lebret, Pedro O. Pinheiro, Ronan Collobert

Idiap Research Institute / EPFL

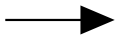
ICML, 9 July 2015

Image Captioning

- Objective: Generate descriptive sentences given a sample image.



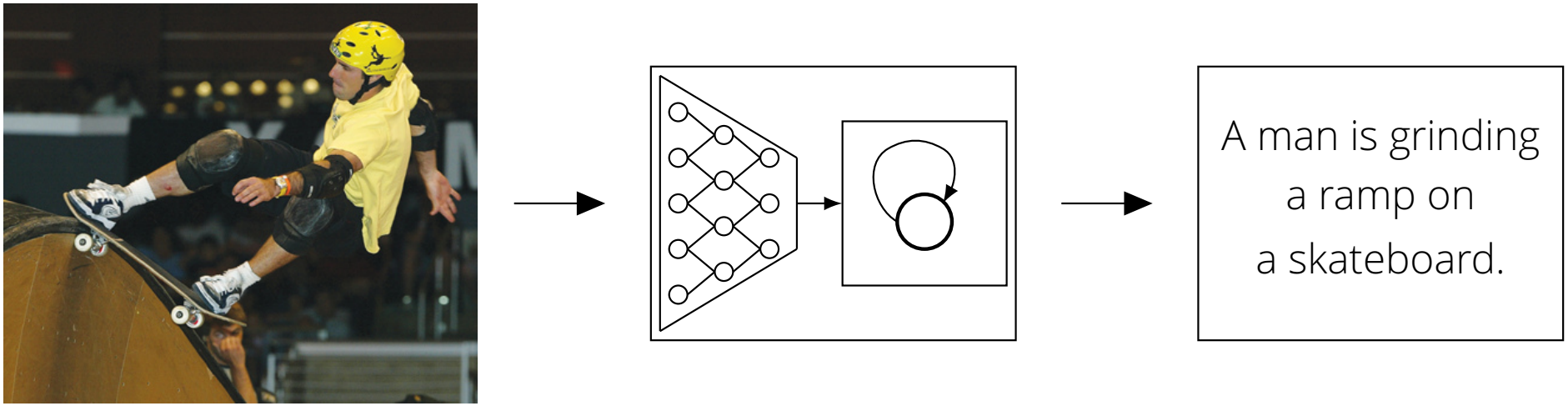
Model



A man is grinding
a ramp on
a skateboard.

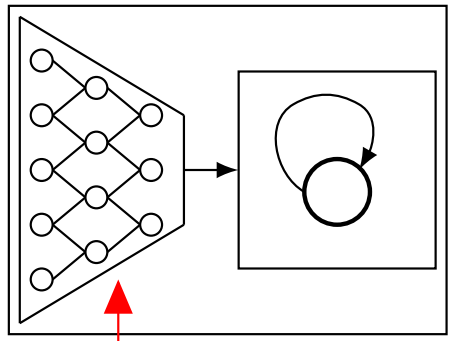
Related Works

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



Related Works

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

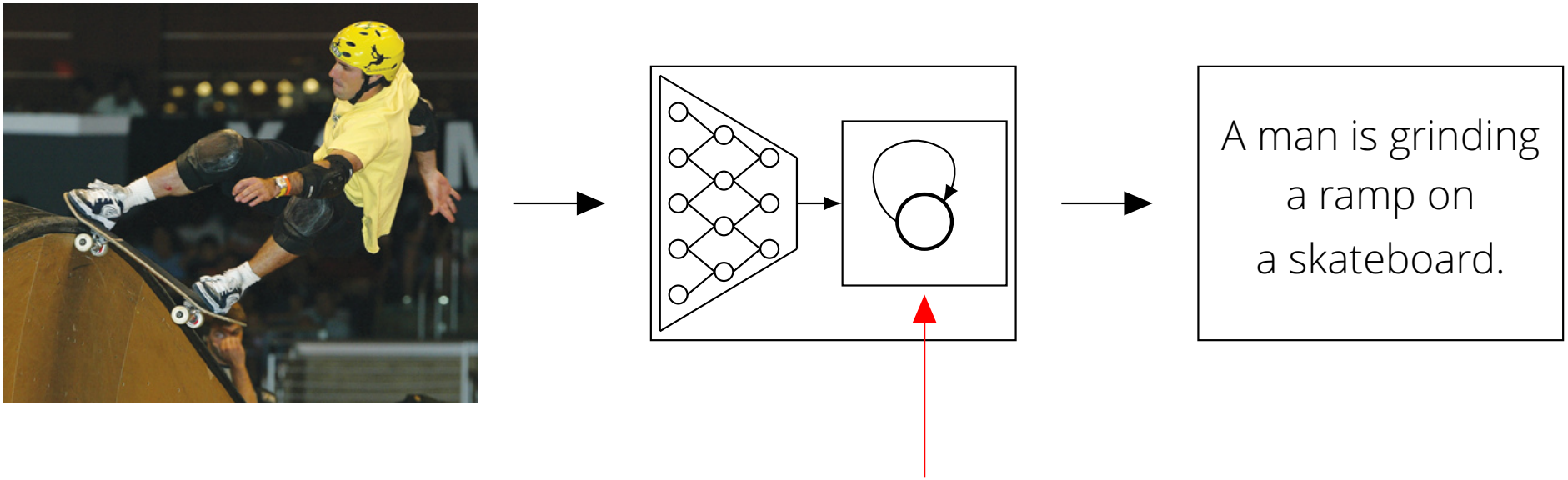


A man is grinding a ramp on a skateboard.

Visual features with Deep CNN

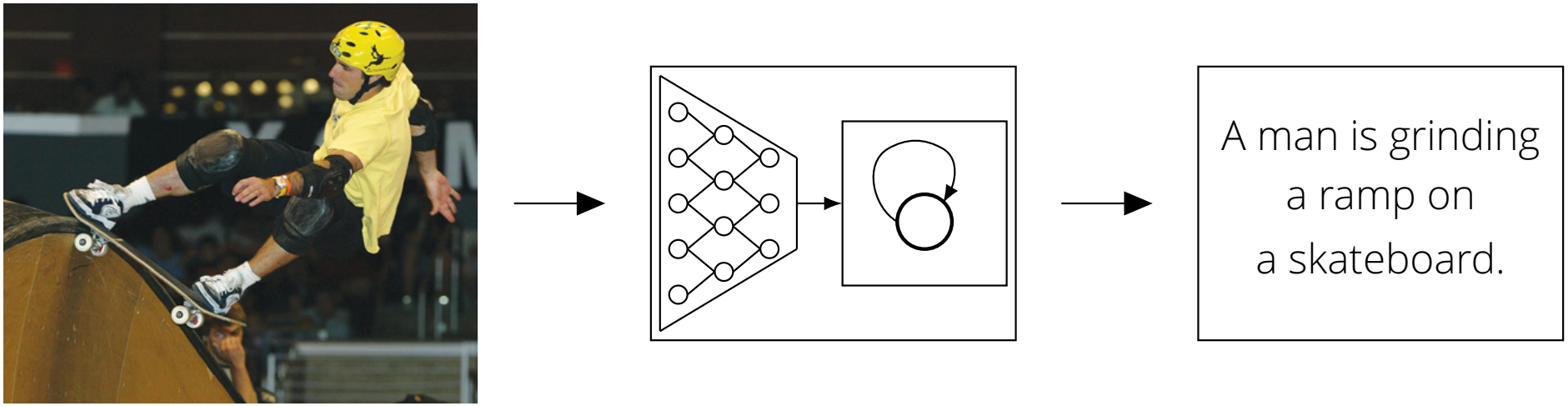
Related Works

- 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 $i \in \mathcal{I}$

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

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
NP VP NP PP NP PP NP
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 (NP) riding (VP) a skateboard (NP) up (PP) the side (NP) of (PP) a wooden ramp (NP)

a man (NP) is grinding (VP) a ramp (NP) on (PP) a skateboard (NP)

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 (NP) riding (VP) a skateboard (NP) up (PP) the side (NP) of (PP) a wooden ramp (NP)

a man (NP) is grinding (VP) a ramp (NP) on (PP) a skateboard (NP)

man (NP) riding on (VP) edge (NP) of (PP) an oval ramp (NP) with (PP) a skate board (NP)

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
NP VP NP PP NP PP NP

a man is grinding a ramp on a skateboard
NP VP NP PP NP

man riding on edge of an oval ramp with a skate board
NP VP NP PP NP PP NP

a man in a helmet skateboarding before an audience
NP PP NP PP NP

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 (NP) riding (VP) a skateboard (NP) up (PP) the side (NP) of (PP) a wooden ramp (NP)

a man (NP) is grinding (VP) a ramp (NP) on (PP) a skateboard (NP)

man (NP) riding on (VP) edge (NP) of (PP) an oval ramp (NP) with (PP) a skate board (NP)

a man (NP) in (PP) a helmet skateboarding (NP) before (PP) an audience (NP)

a man (NP) on (PP) a skateboard (NP) is doing (VP) a trick (NP)

→ 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 (NP) riding (VP) a skateboard (NP) up (PP) the side (NP) of (PP) a wooden ramp (NP)

a man (NP) is grinding (VP) a ramp (NP) on (PP) a skateboard (NP)

man (NP) riding on (VP) edge (NP) of (PP) an oval ramp (NP) with (PP) a skate board (NP)

a man (NP) in (PP) a helmet skateboarding (NP) before (PP) an audience (NP)

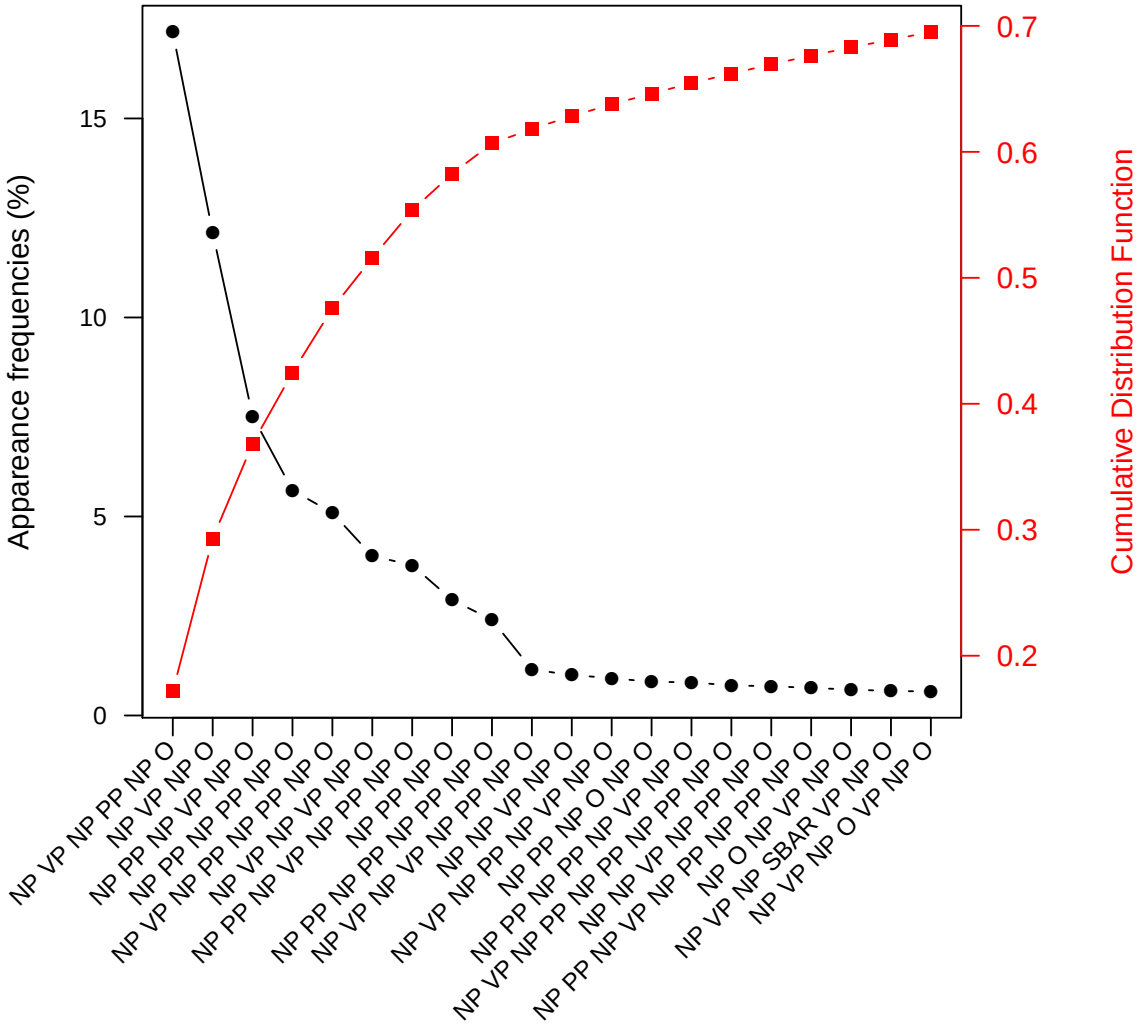
a man (NP) on (PP) a skateboard (NP) is doing (VP) a trick (NP)

- Noun phrases (NP) → Key elements in images.
- 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.

Phrase-based Model for Image Descriptions

Our approach:

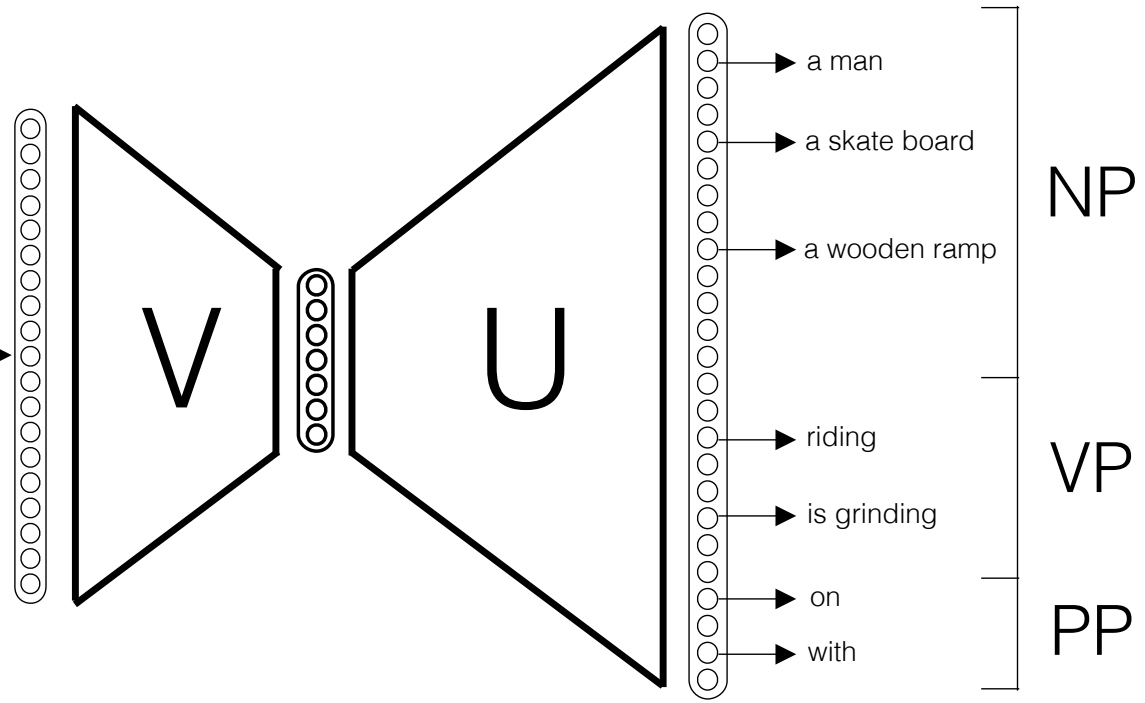
1. A bilinear model that learns a metric between an image and phrases used to describe it.
2. Sentences generated using a simple language model based on caption syntax statistics.

A Bilinear Model $U^T V$

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

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

A man in a helment 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.



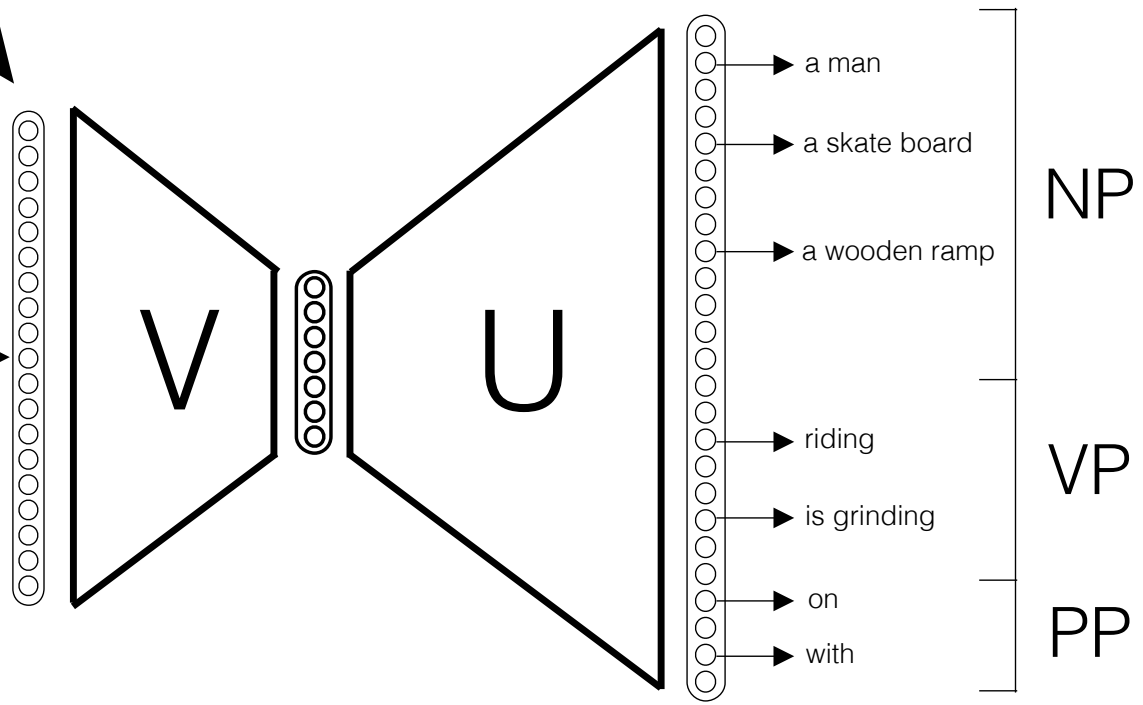
A Bilinear Model $U^T V$

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

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

pre-trained CNN representation $\mathbf{z}_i \in \mathbb{R}^n$

A man in a helment 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.



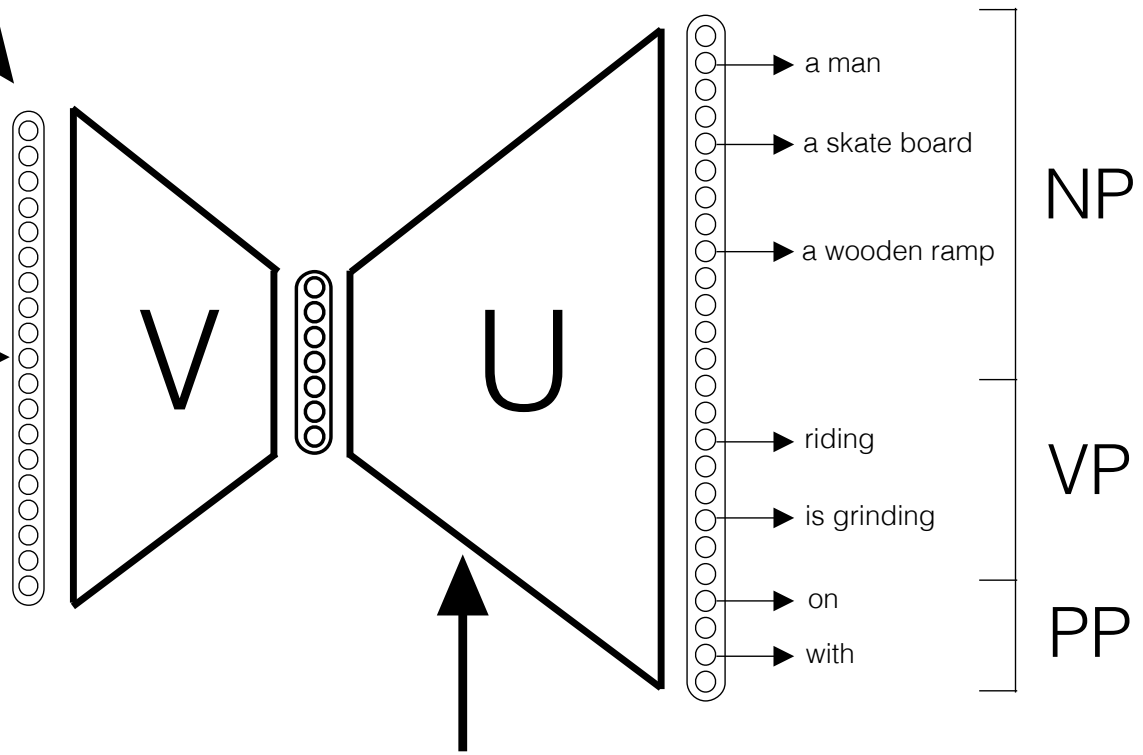
A Bilinear Model $U^T V$

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

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

pre-trained CNN representation $\mathbf{z}_i \in \mathbb{R}^n$

A man in a helment 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.



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

$$\mathbf{u}_c = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{w_k}$$

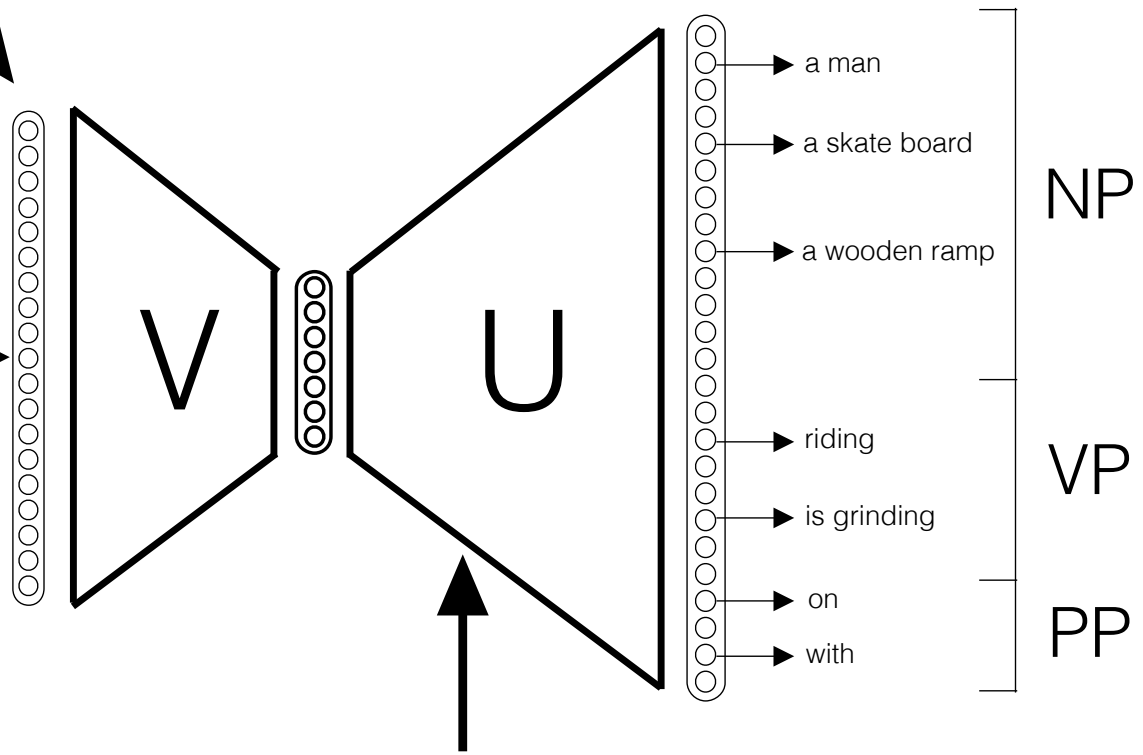
A Bilinear Model $U^T V$

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

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

pre-trained CNN representation $\mathbf{z}_i \in \mathbb{R}^n$

A man in a helment 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.



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

$$\mathbf{u}_c = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{w_k}$$

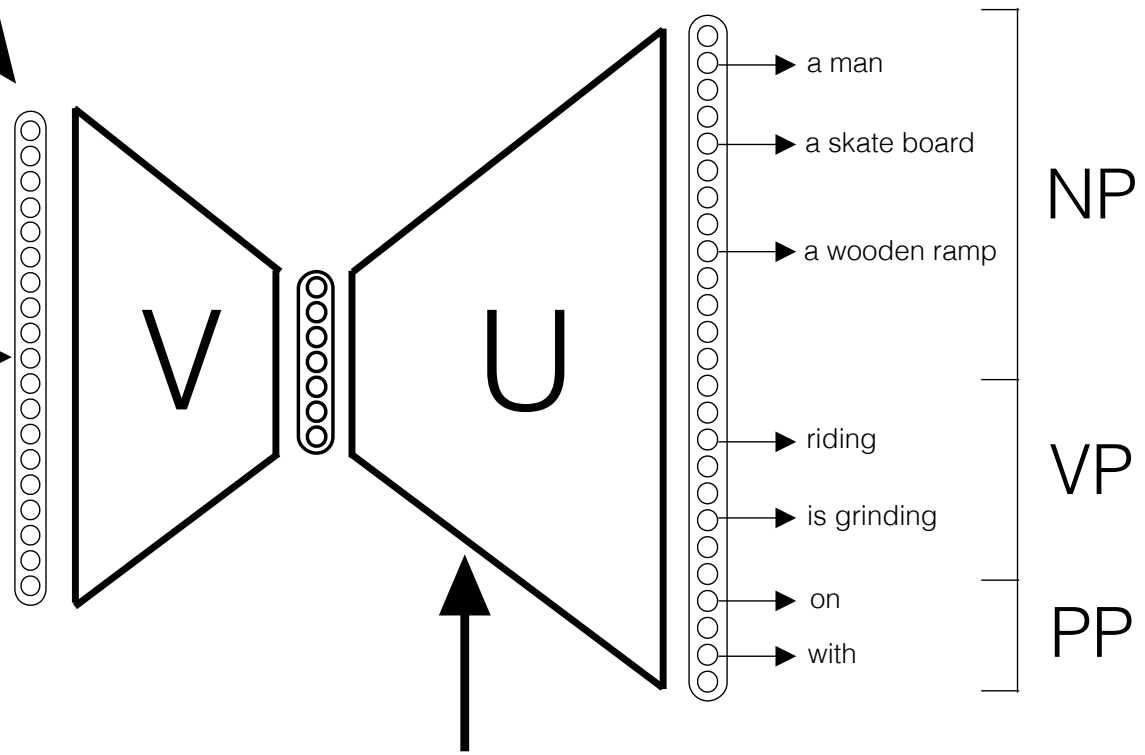
A Bilinear Model $U^T V$

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

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

pre-trained CNN representation $\mathbf{z}_i \in \mathbb{R}^n$

A man in a helment 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.



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

$$\mathbf{u}_c = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{w_k}$$

A Bilinear Model $U^T V$

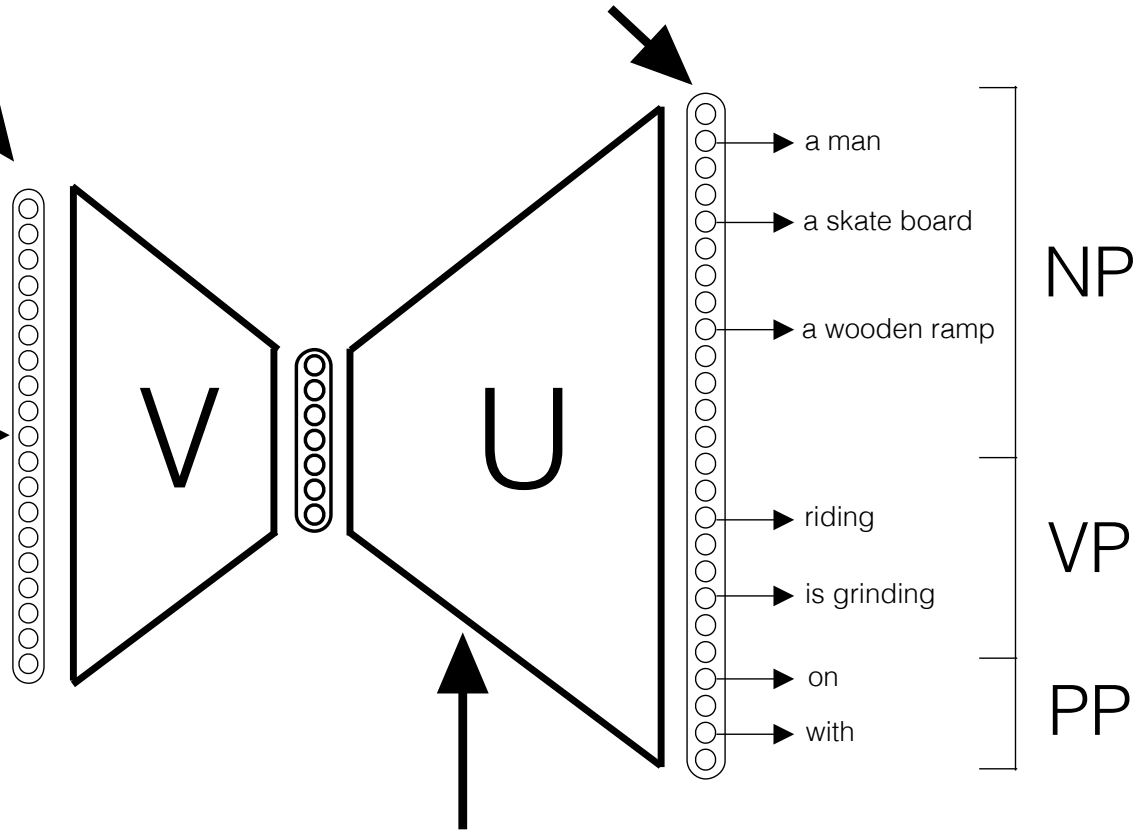
$\mathcal{I}$ = set of training images
$\mathcal{C}$ = set of all phrases used to describe $\mathcal{I}$

$U = (\mathbf{u}_{c_1}, \ldots, \mathbf{u}_{c_{|\mathcal{C}|}}) \in \mathbb{R}^{m \times |\mathcal{C}|}$
$V \in \mathbb{R}^{m \times n}$
} trainable parameters θ

score between the image i and a phrase c: $f_\theta(c, i) = \mathbf{u}_c^T V \mathbf{z}_i$

pre-trained CNN representation $\mathbf{z}_i \in \mathbb{R}^n$

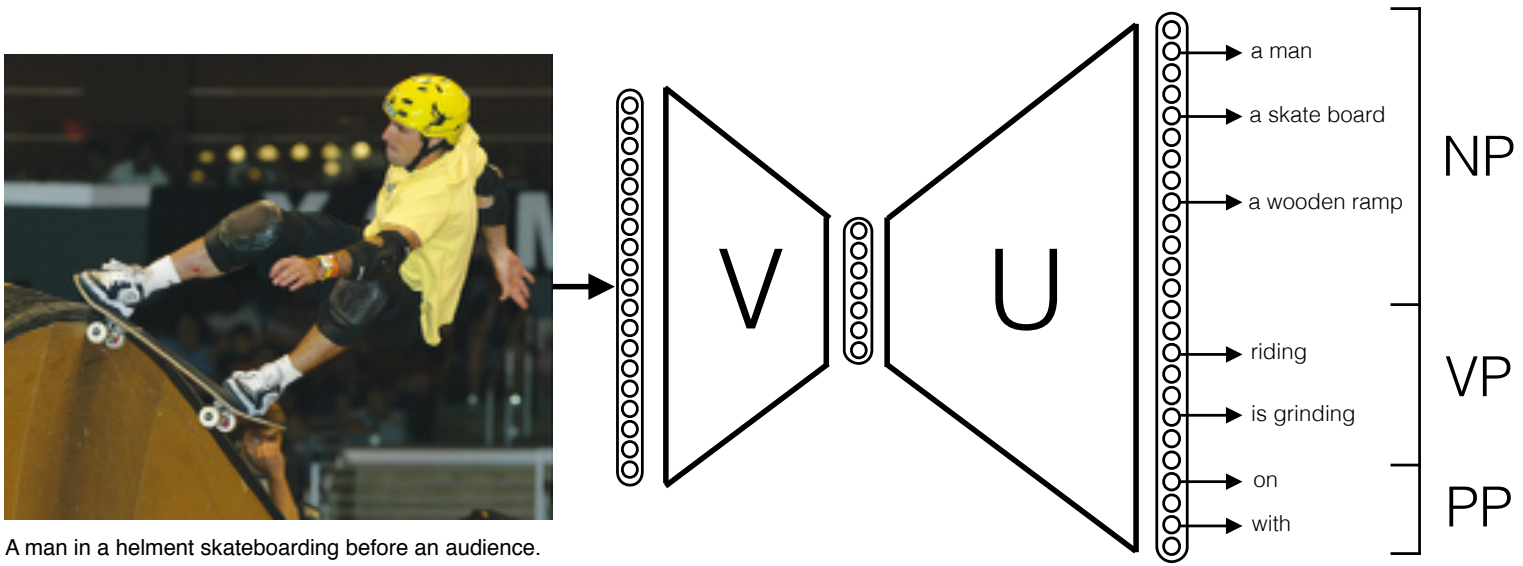
A man in a helment skateboarding before an audience.
Man riding on edge of an oval ramp with a skate board.
A man riding a skateboard up the side of a wooden ramp.
A man on a skateboard is doing a trick.
A man is grinding a ramp on a skateboard.



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

$$\mathbf{u}_c = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_{w_k}$$

A Bilinear Model $U^T V$



A man in a helment 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.

Training with negative sampling by minimizing this logistic loss function w.r.t. θ:

$$\theta \mapsto \sum_{i \in \mathcal{I}} \sum_{c_j \in \mathcal{C}^i} \left(\log \left(1 + e^{-\mathrm{u}_{c_j}^T V \mathrm{z}_i} \right) + \sum_{c_k \in \mathcal{C}^-} \log \left(1 + e^{+\mathrm{u}_{c_k}^T V \mathrm{z}_i} \right) \right)$$

$\rightarrow$ Stochastic gradient descent, new set of negative phrases $\mathcal{C}^-$ 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:

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

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

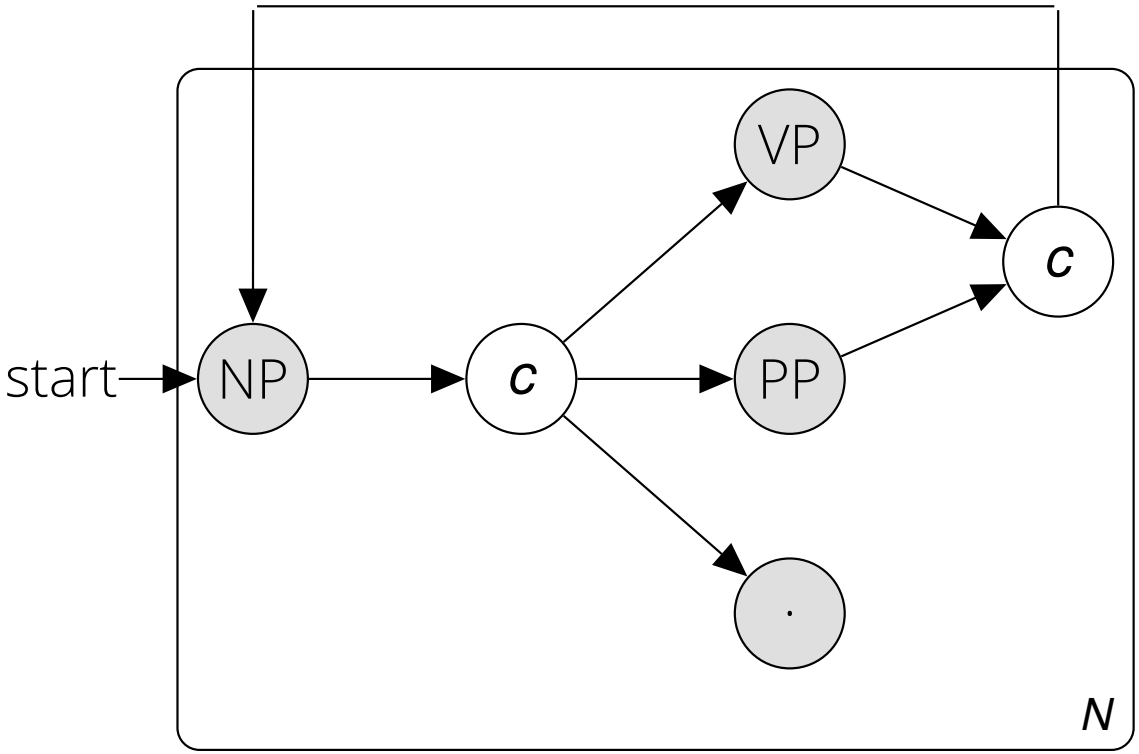
- Prior knowledge on chunking tags $t \in \{NP, VP, PP\}$:

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

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

Sentence Decoding

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



$N \in \{2, 3, 4\}$

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

$\rightarrow$ 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: $f_\theta(c, i) = \mathbf{u}_c^T V \mathbf{z}_i$.
- Averaging phrase scores $f_\theta(c_j, i) \, \forall c_j \in s$:

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

$\rightarrow$ 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 θ:
  - $V \in \mathbb{R}^{400 \times 4096} \rightarrow$ initialized randomly.
  - $U \in \mathbb{R}^{400 \times |\mathcal{C}|} \rightarrow$ 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

Captioning Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Human agreement	0.68	0.45	0.30	0.20
CNN/RNN based models				
Mao *et al.*	0.67	0.49	0.35	0.25
Karpathy & Fei-Fei	0.63	0.45	0.32	0.23
Vinyals *et al.*	0.67	-	-	-
Donahue *et al.*	0.63	0.44	0.30	0.21
Our model	0.73	0.50	0.34	0.23

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

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

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

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

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

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

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

Failure example

a flock of geese are walking in a parking lot

Failure example

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

Failure example

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

PHRASES		NEAREST NEIGHBORS	
	#	BEFORE	AFTER
A GREY CAT	1	A GREY DOG	A GRAY CAT
	2	A GREY AND BLACK CAT	A GREY AND BLACK CAT
	3	A GRAY CAT	A BROWN CAT
	4	A GREY ELEPHANT	A GREY AND WHITE CAT
	10	A YELLOW CAT	GREY AND WHITE CAT
HOME PLATE	1	A HOME PLATE	A HOME PLATE
	4	A PLATE	HOME BASE
	6	ANOTHER PLATE	THE PITCH
	9	A RED PLATE	THE BATTER
	10	A DINNER PLATE	A BASEBALL PITCH
A HALF PIPE	1	A PIPE	A PIPE
	2	A HALF	THE RAMP
	5	A SMALL CLOCK	A HAND RAIL
	9	A LARGE CLOCK	A SKATE BOARD RAMP
	10	A SMALL PLATE	AN EMPTY POOL

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