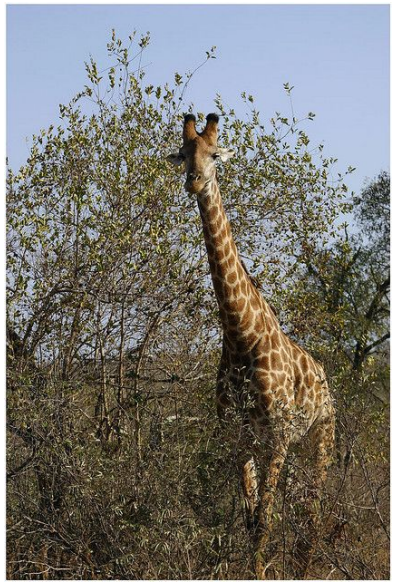
Neural Networks and Fuzzy Logic - BITS F312

Text-guided Attention Model for Image Captioning

Submitted By:
Vikram Nitin (2015A7PS0042P)
Rohan Saxena (2015A7TS0017P)
Naveen Venkat (2015A7PS0078P)

Introduction

- Image captioning produces textual description of the image



“A giraffe is walking in a grassy field”

“A donut is kept on the table”

“A pizza is kept on a plate”

Attention based method

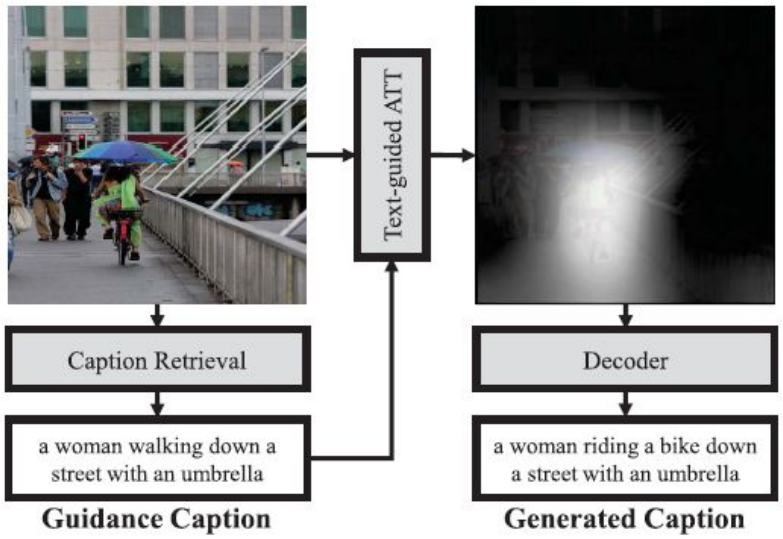
- **Focus** only on the **relevant locations** while generating the captions
- **Blur** out **other regions**

"A **woman** riding a **bike** down a **street** with an **umbrella**"

Whiteness: degree of importance of features

Text Guided Attention

Use “guidance captions” which can be retrieved from **similar images**, to **guide the attention** to specific regions of the image



Formulation of Problem & Objectives

- Extract features from image that helps captioning
- Use **guidance captions** to improve the accuracy of captions
- Encode these two, pass through an attention model, give to decoder
- Decoder will generate captions
- Evaluate the accuracy

Source of Data

MS-COCO dataset (2014) :

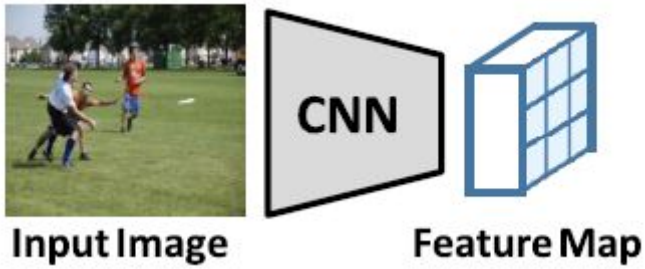
- **123,287 images (82,783 training images and 40,504 validation images)**
- **cocodataset.org**

Libraries Used

- Python
  - Keras
  - Tensorflow (backend)
  - CIDEr (adapted form https://github.com/vrama91/cider)
  - Skip-Thought Vectors (Tensorflow provided: https://github.com/tensorflow/models/tree/master/research/skip_thoughts)
  - NLTK

Feature Extraction

- Extract **features** out of the image using convolution operation
- CNNs used for analysis: VGG-FCN, Resnet-101
- CNNs are fine tuned on MS-COCO dataset



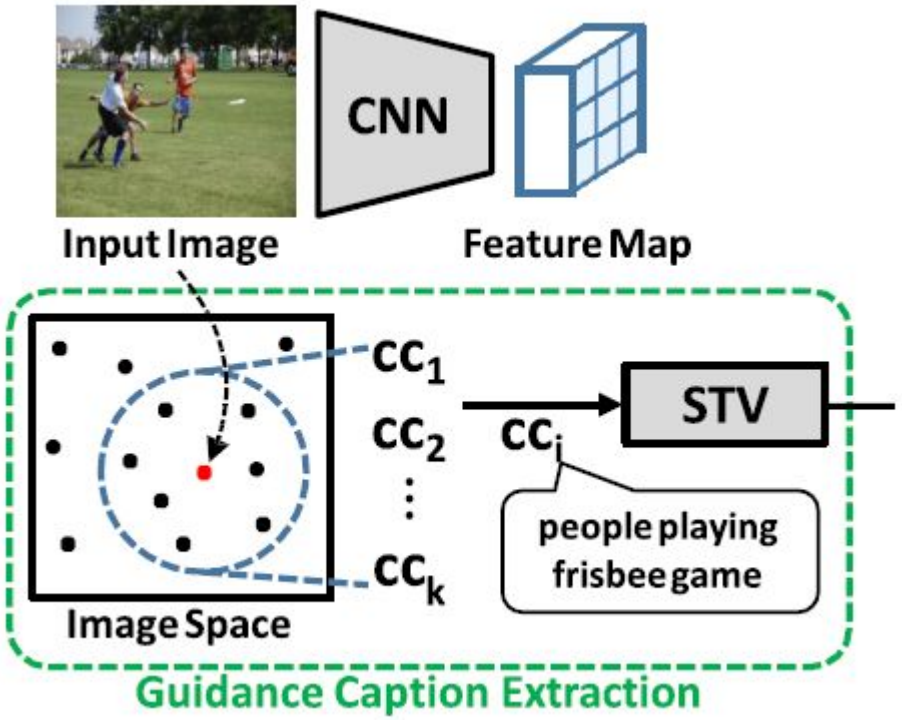
Guidance Caption

- Nearest neighbors approach
- Visually similar images tend to share the salient objects and events that are often described in their captions

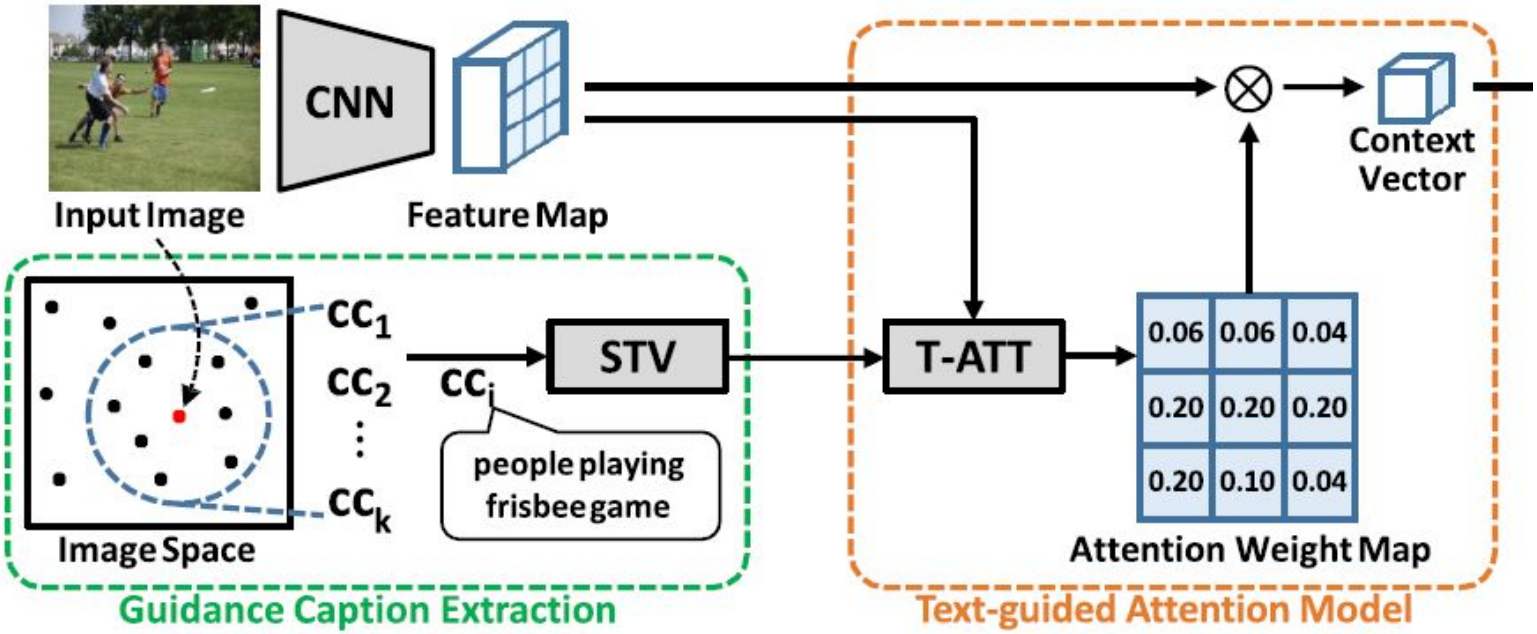
$$s_i = \frac{1}{|C_{\text{NN}}| - 1} \sum_{c' \in C_{\text{NN}} \setminus \{c_i\}} \text{Sim}(c_i, c'), \qquad (3)$$

C_{NN} : set of all closest captions

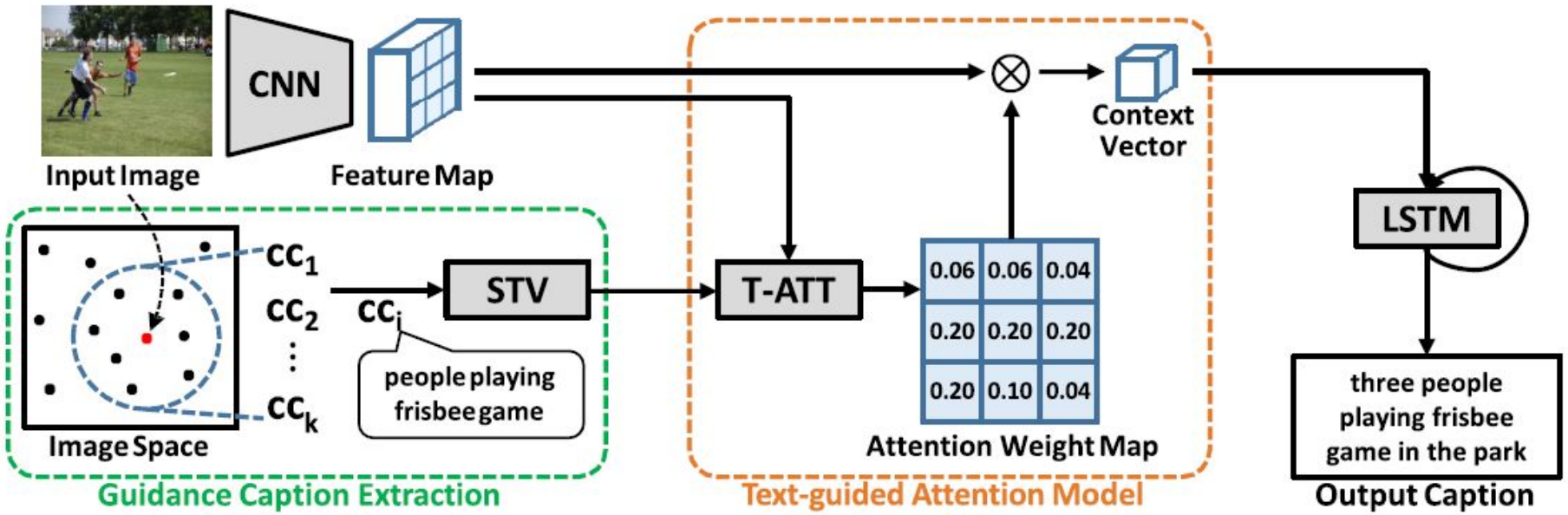
Sim(ci, c') : CIDEr similarity score



Text-guided Model



Architecture of the Network



Performance on MS-COCO Dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
NIC (Vinyals et al. 2015)$^{\diamond}$	–	–	–	0.277	0.237	0.855
m-RNN (Mao et al. 2015b)	0.686	0.511	0.375	0.280	0.228	0.842
LRCN (Donahue et al. 2015)	0.669	0.489	0.349	0.249	–	–
MSR (Fang et al. 2015)†	–	–	–	0.257	0.236	–
Soft-ATT (Xu et al. 2015)	0.707	0.492	0.344	0.243	0.239	–
Hard-ATT (Xu et al. 2015)	0.718	0.504	0.357	0.250	0.230	–
Semantic ATT (You et al. 2016)$^{\dagger\diamond}$	0.709	0.537	0.402	0.304	0.243	–
Attribute-LSTM (Wu et al. 2016)†	**0.740**	0.560	0.420	0.310	**0.260**	0.940
mCC	0.670	0.486	0.345	0.244	0.225	0.791
Ours-Uniform (VGG-FCN)	0.733	0.563	0.420	0.313	0.249	0.968
Ours-ATT-mCC (VGG-FCN)	0.732	0.563	0.421	0.313	0.250	0.972
Ours-ATT-kCC (VGG-FCN)	0.735	**0.566**	**0.424**	**0.316**	0.251	**0.982**
Ours-Uniform (ResNet)	0.741	0.572	0.429	0.319	0.253	0.996
Ours-ATT-mCC (ResNet)	0.743	0.575	0.432	0.323	0.255	1.010
Ours-ATT-kCC (ResNet)	**0.749**	**0.581**	**0.437**	**0.326**	0.257	**1.024**

red = best, blue = second best

Performance on COCO image captioning challenge

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
NIC (Vinyals et al. 2015)†◇	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	**0.254**	**0.346**	**0.943**	0.946
m-RNN (Mao et al. 2015b)	0.716	0.890	0.545	0.798	0.404	0.687	0.299	0.575	0.242	0.325	0.917	0.935
LRCN (Donahue et al. 2015)†	0.718	0.895	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.921	0.934
MSR (Fang et al. 2015)†	0.715	**0.907**	0.543	**0.819**	0.407	**0.710**	0.308	0.601	0.248	0.339	0.931	0.937
Hard-ATT (Xu et al. 2015)	0.705	0.881	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.865	0.893
Semantic ATT (You et al. 2016)†◇	**0.731**	0.900	**0.565**	0.815	**0.424**	0.709	**0.316**	**0.599**	0.250	0.335	**0.943**	**0.958**
Attribute-LSTM (Wu et al. 2016)†	0.725	0.892	0.556	0.803	0.414	0.694	0.306	0.582	0.246	0.329	0.911	0.924
Ours-ATT-mCC (VGG-FCN)	0.727	0.901	0.558	0.812	0.414	0.701	0.305	0.585	0.247	0.330	0.930	0.942
Ours-ATT-kCC (VGG-FCN)	**0.731**	0.902	0.560	0.814	0.416	0.703	0.307	0.587	0.248	0.331	0.938	0.951
Ours-ATT-mCC (ResNet)	0.739	0.913	0.570	0.828	0.427	0.719	0.318	0.603	0.253	0.339	0.972	0.988
Ours-ATT-kCC (ResNet)	**0.743**	**0.915**	**0.575**	**0.832**	**0.431**	**0.722**	**0.321**	**0.607**	**0.255**	0.341	**0.987**	**1.001**

red = best, blue = second best

a sandwich sitting on top of a white plate
a sandwich cut in half on a plate
a sandwich is on a plate on a table

a man **sitting on a couch** on a cell phone
a man **holding** a cell phone **in his hand**
a man **sitting at a table** with a cell phone

a group of people standing **around a beach**
a group of people standing **on a beach with surfboards**
a group of people standing **on top of a beach**

a piece of cake sitting **on top of a plate**
a piece of cake **on a plate with a fork**
a piece of chocolate cake **on a white plate**

a white passenger bus is parked on a street
a man standing next to a bus on a road
a bus is parked on the side of the road

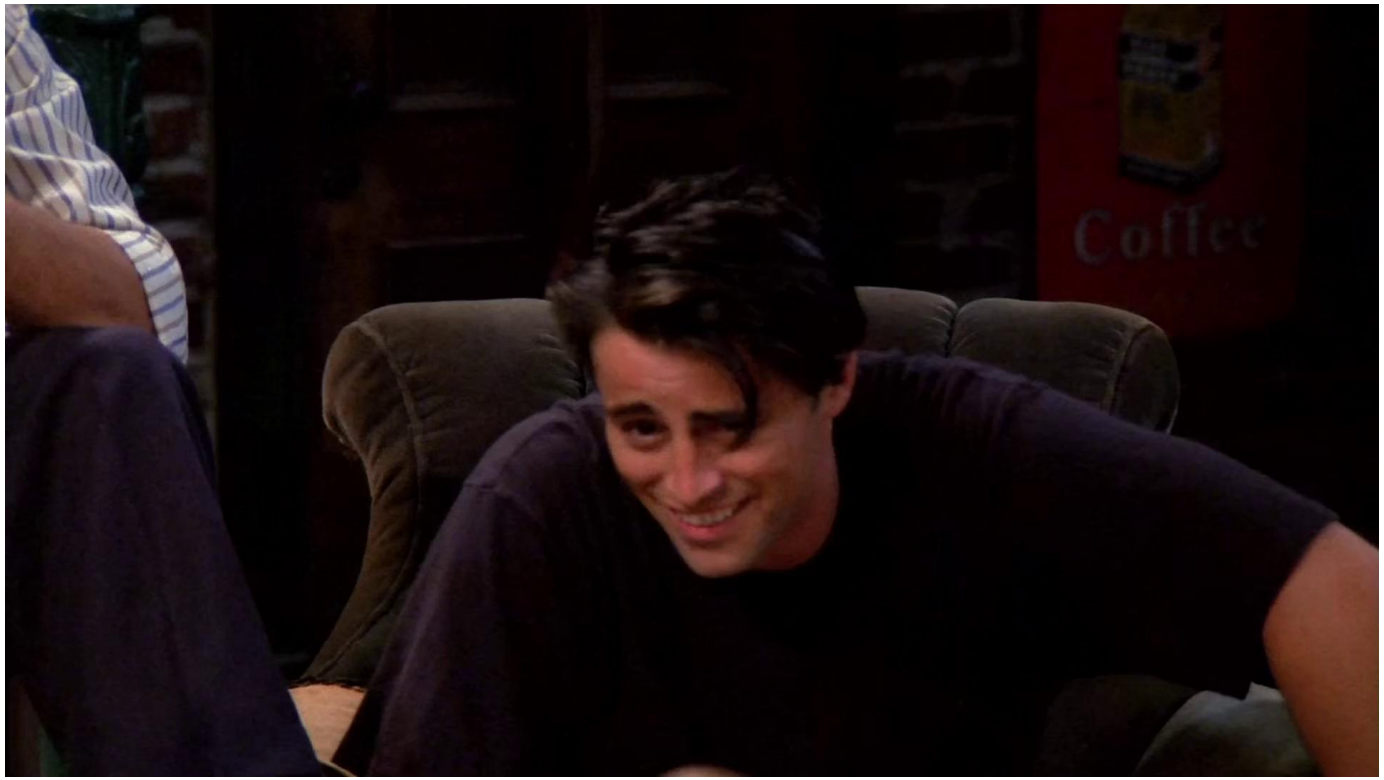
a boat that is floating in the water
a group of birds sitting on top of a boat
two birds are sitting on a dock in the water

a toilet lying on its side on a sidewalk
an old refrigerator sitting in front of a wall
a old fashioned wall with **a clock** on it

a pile of lots of broccoli on a table
a close up of a bunch of broccoli
a close up of broccoli on a table

a man riding a motorcycle on the street
a group of people riding bikes down a street
a man riding a bike down a street

two women and a young girl pose in a kitchen



a man sitting in a chair talking on a cell phone

a man in a suit and bow tie looking at the camera

Conclusion

Text guided method provides “exemplar-based” learning method.

Guidance captions highlight the relevant regions

A set of consensus captions can be exploited to reduce noise