A Hierarchical Approach for Generating Descriptive Image Paragraphs

Jonathan Krause, Justin Johnson, Ranjay Krishna, Li Fei-Fei

Presented by Tianyang Liu
Feb 1, 2017

IMAGE CAPTIONING

- One sentence description
  - A great amount of detail is left out
- Multi-sentence description (dense captioning)
  - Solves the lack of detail problem, but sentences are not coherent
- **Paragraph description**

Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

RELATED WORK #1

- **Baby talk: Understanding and generating image descriptions.** [G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011]

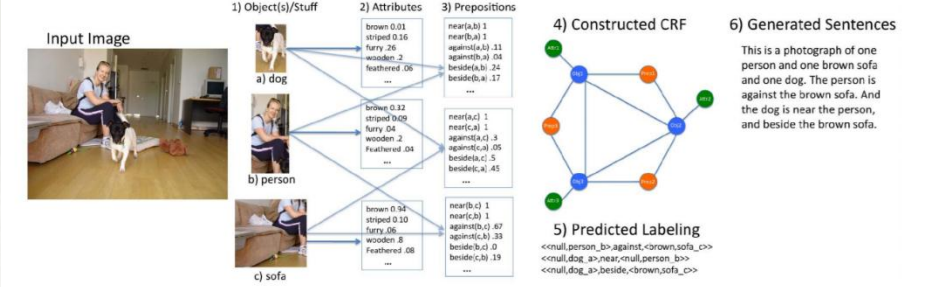


Fig. 2. System flow for an example image: (1) Object and stuff detectors find candidate objects, (2) each candidate region is processed by a set of attribute classifiers, (3) each pair of candidate regions is processed by prepositional relationship functions, (4) a CRF is constructed that incorporates the unary image potentials computed by 1-3 and higher order text-based potentials computed from large document corpora, (5) a labeling of the graph is predicted, and (6) sentences are generated.

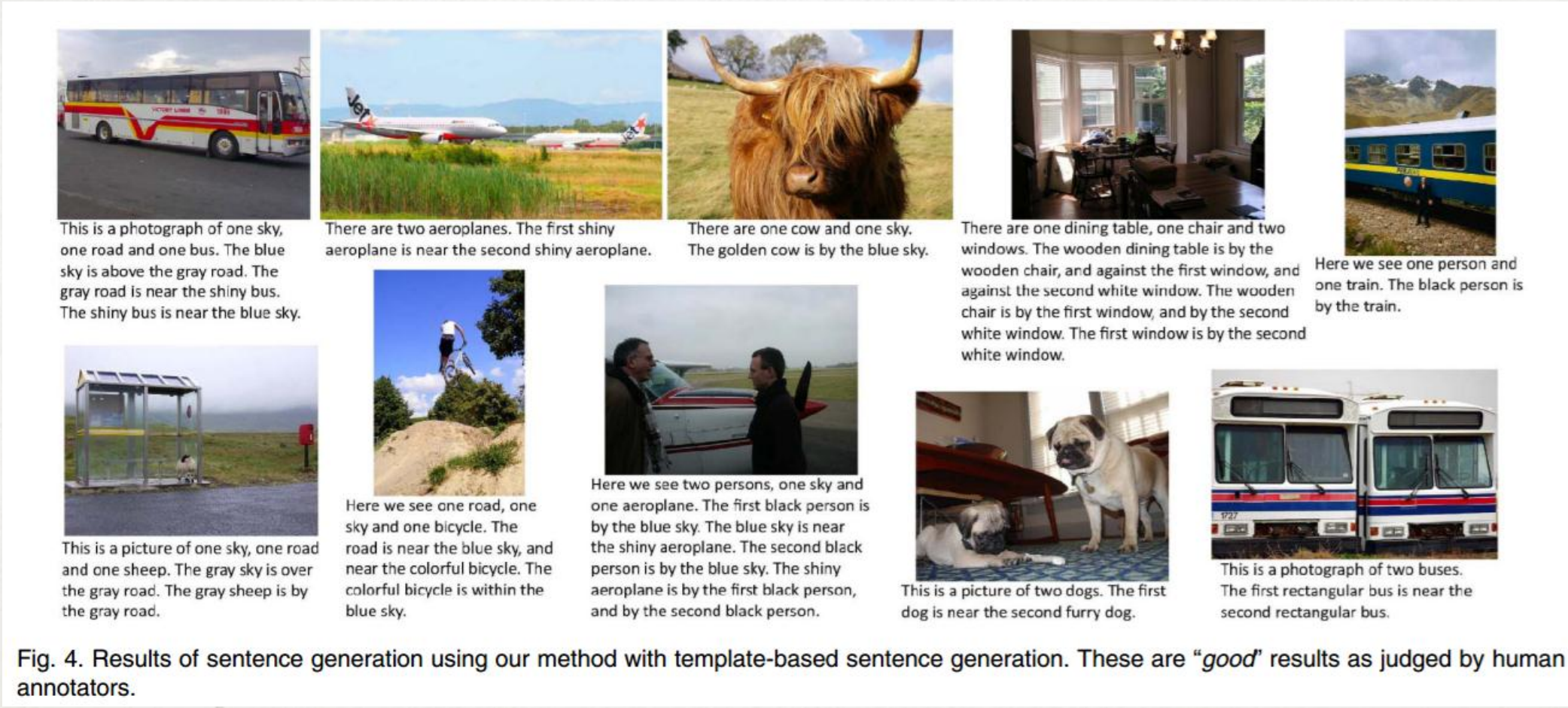
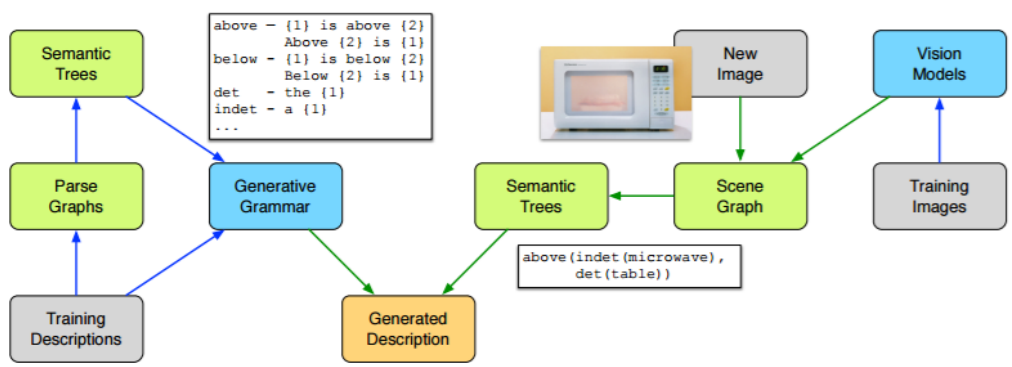


Fig. 4. Results of sentence generation using our method with template-based sentence generation. These are "*good*" results as judged by human annotators.

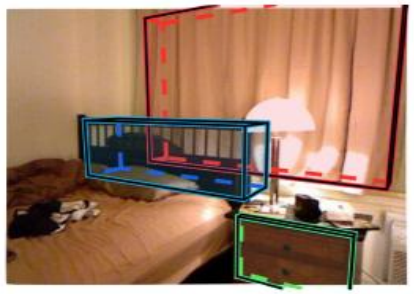
Figures from G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In CVPR, 2011

RELATED WORK #2

- **Generating Multi-sentence Natural Language Descriptions of Indoor Scenes**
  [Dahua Lin, Sanja Fidler, Chen Kong, Raquel Urtasun. 2015]

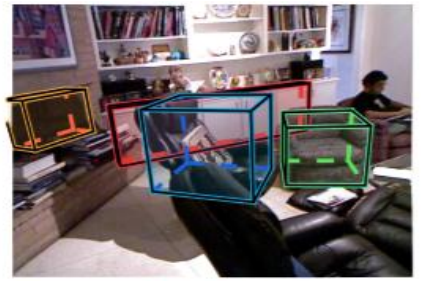


There is a brown bed in the bedroom. The bed is in front of a headboard. Near the bed is a blinds. We can see a brown curtain near the blinds. There is a chest near the headboard.



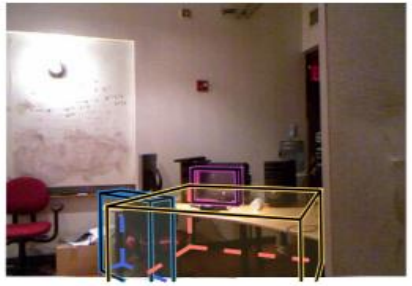
A wooden curtain is in the bedroom. The curtain is on top of a wooden headboard. We can see a chest in front of the curtain. The headboard is near the chest.

In the kitchen, there is a refrigerator. A green cabinet is near a gray oven. Near the refrigerator is the cabinet. We can see a microwave near the cabinet. The oven is behind the refrigerator.

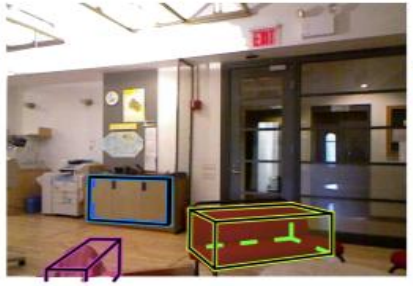


There is a sofa in the living room. Behind the sofa is a white cabinet. We can see a black chair in front of the cabinet. There is a mantel near the chair.

In the office, there is a board. We can see a cabinet in front of the board. We can see a monitor near the board. In the office, there is a table.



In the living room, there is a monitor. The monitor is behind a chair. We can see the monitor on top of a table. There is the table near the monitor. The chair is near the table.

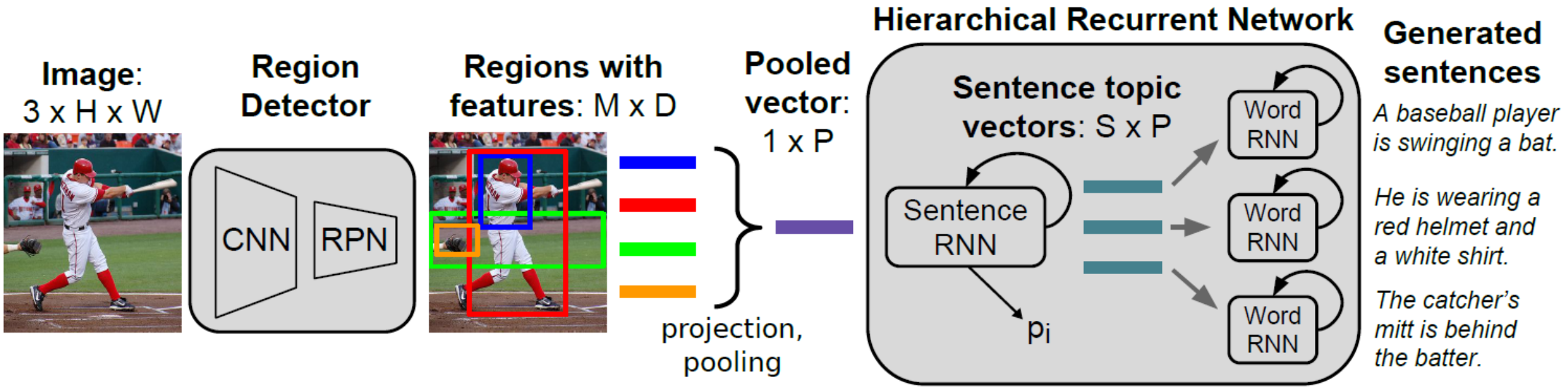


In the kitchen, there is a chair. A cabinet is behind a sofa. The sofa is near the chair.

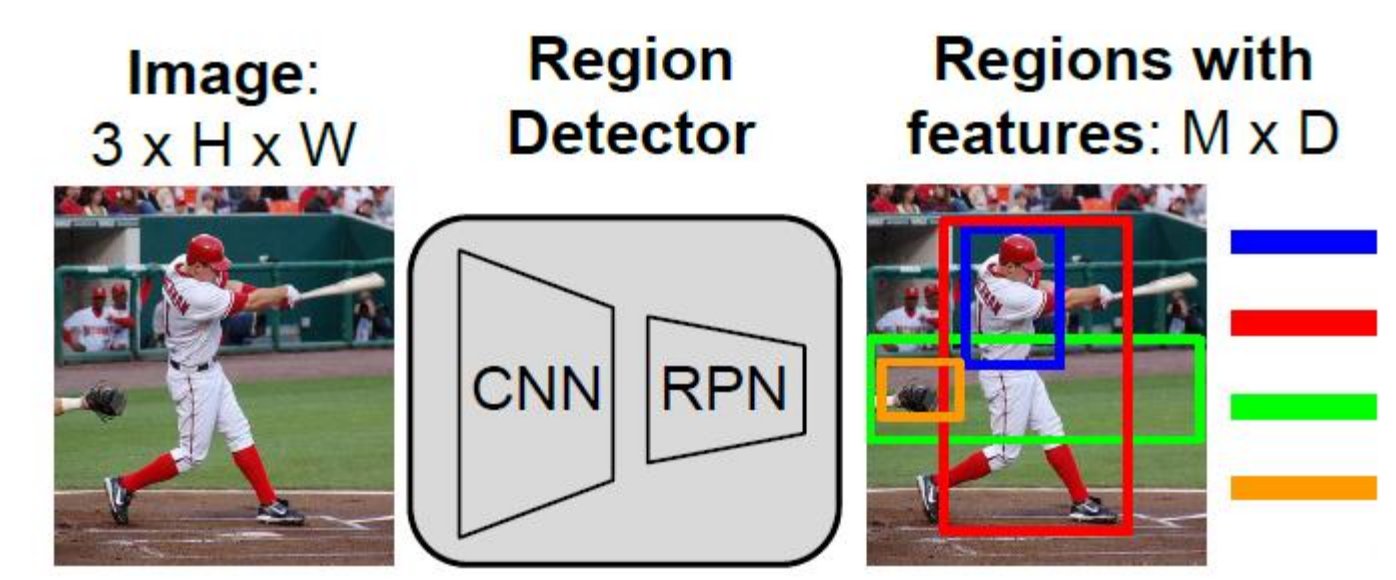
There is a white counter in the kitchen. The counter is near a white cabinet. Near a refrigerator is the cabinet. We can see a green microwav on the right of the cabinet. The refrigerator is near a shelf.

Figures from Generating Multi-sentence Natural Language Descriptions of Indoor Scenes, Dahua Lin, Sanja Fidler, Chen Kong, Raquel Urtasun. 2015

OVERVIEW OF MODEL



REGION DETECTOR



- The image is first run through a pre-trained CNN (16-layer VGG) to extract CNN features
- Given the features, the Region Proposal Network will output the features of *M* most confident regions
- Details of RPN on next slide

REGION PROPOSAL NETWORK

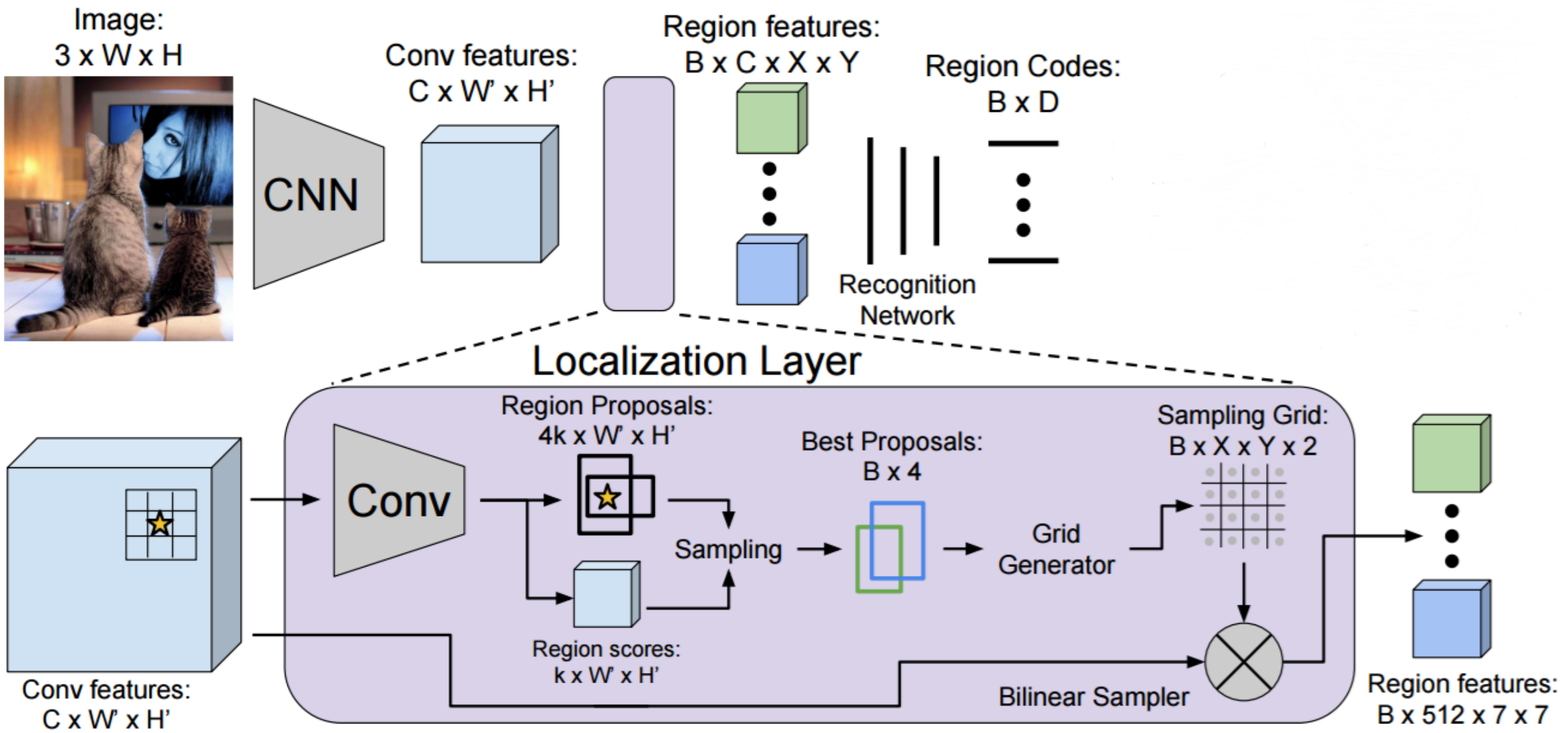
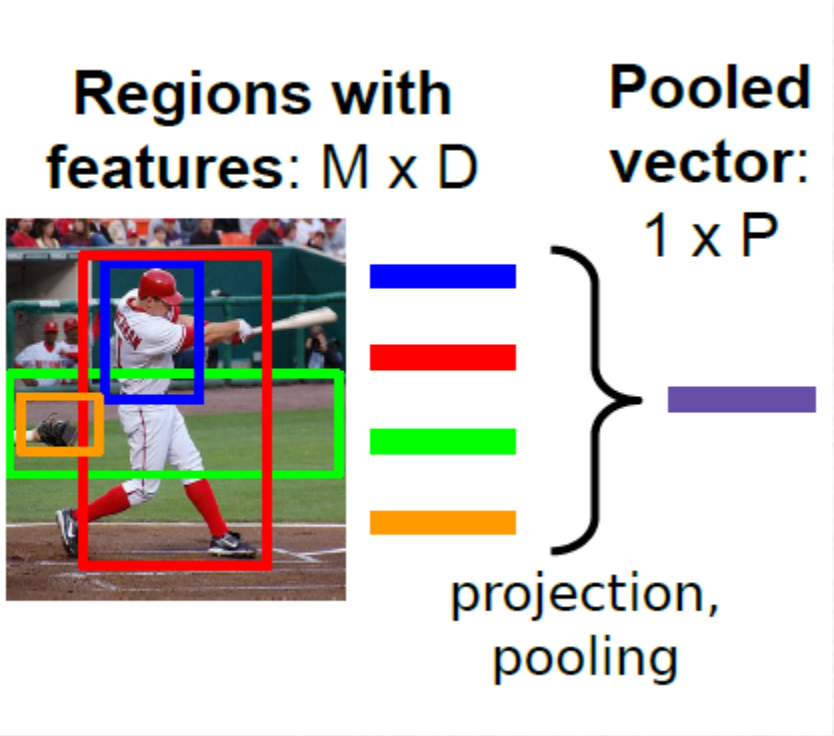


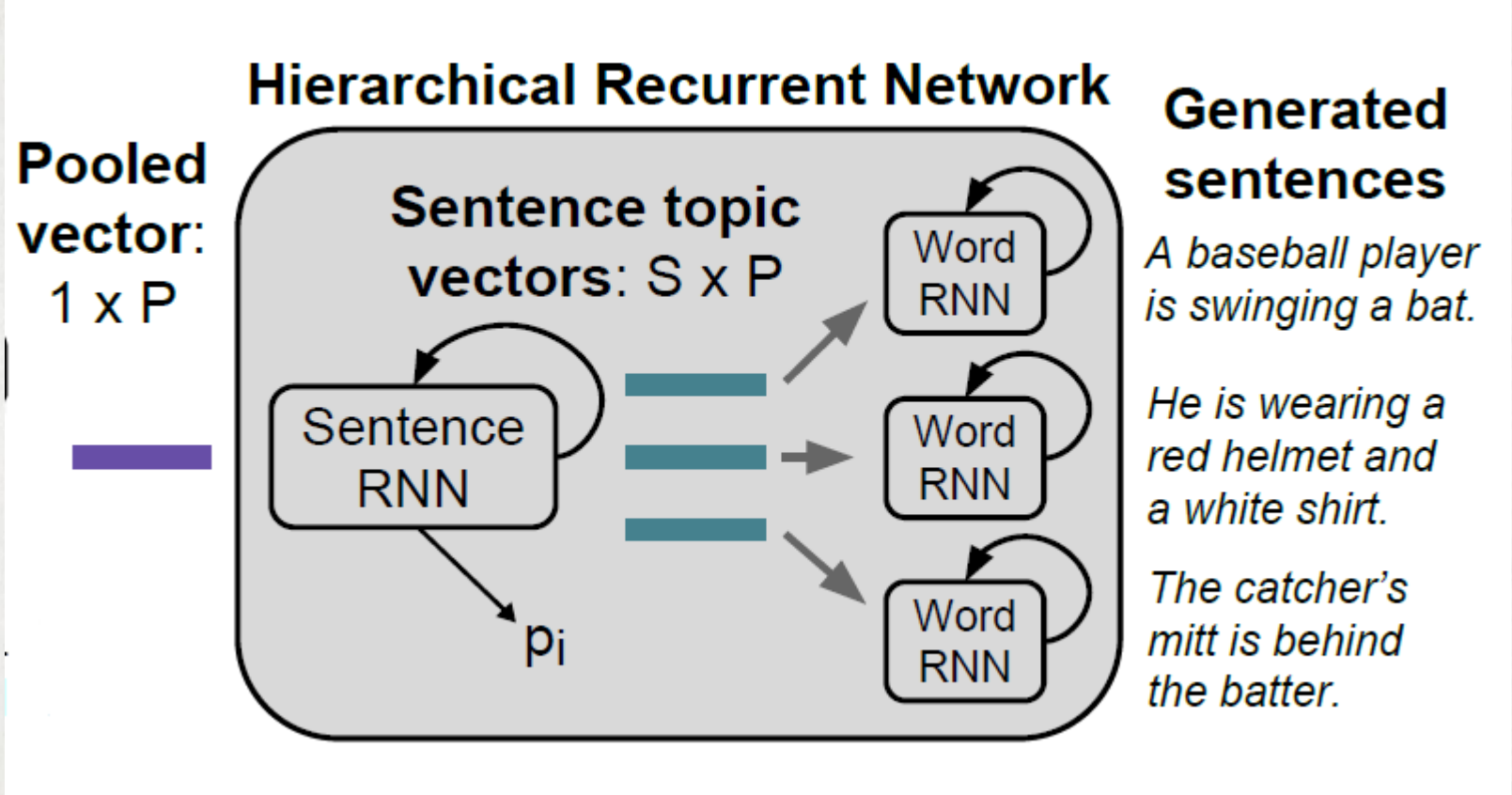
Figure from J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In CVPR, 2016.

REGION POOLING



- Given a set of vectors v_1, …, $v_M \in R^D$, each describing the features of a different region in the input image
- Will learn a projection matrix $W_{pool} \in R^{P \times D}$ and bias $b_{pool} \in R^P$ to create a single pooled vector
- Take the maximum at each element
- The result pooled vector is fed into the hierarchical recurrent neural network language model

HIERARCHICAL RECURRENT NEURAL NETWORK



Includes 2 parts:

- **Sentence RNN**
- **Word RNN**

SENTENCE RNN

Single-layer LSTM with hidden size H = 512

2 Tasks:

- Decide the number of sentences S that should be in the generated paragraph
- Produce a *P*-dimensional topic vector for each of these sentences.

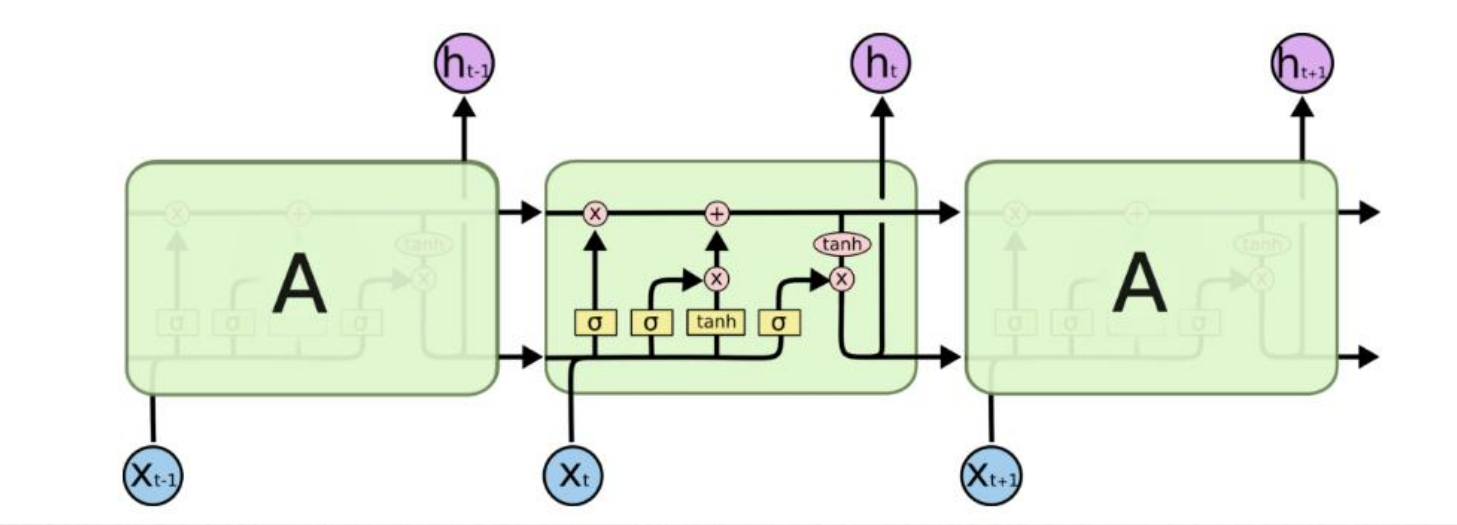
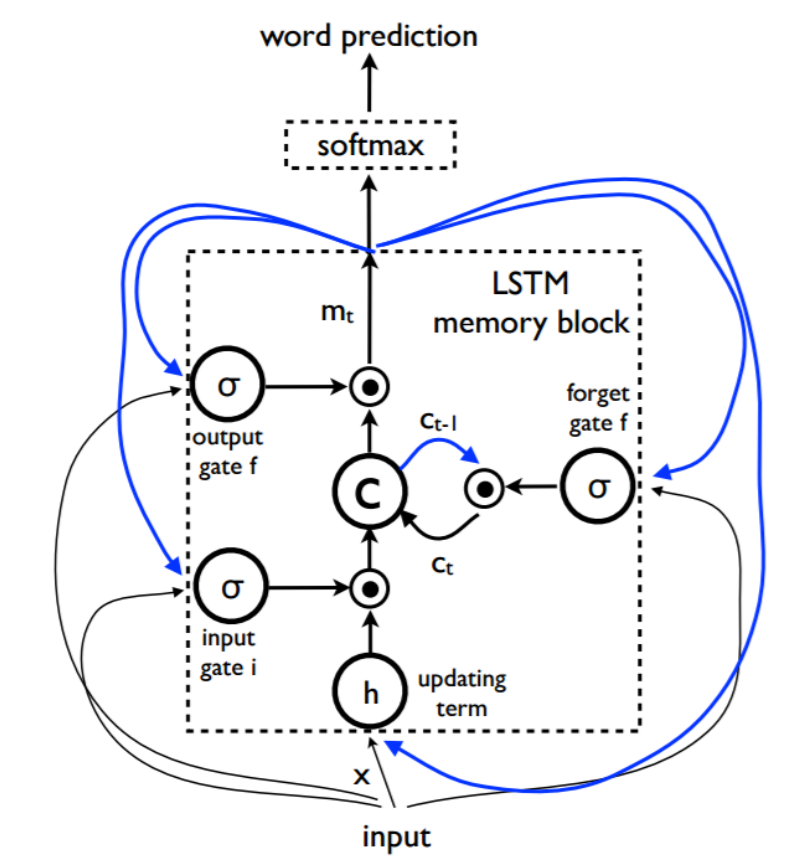
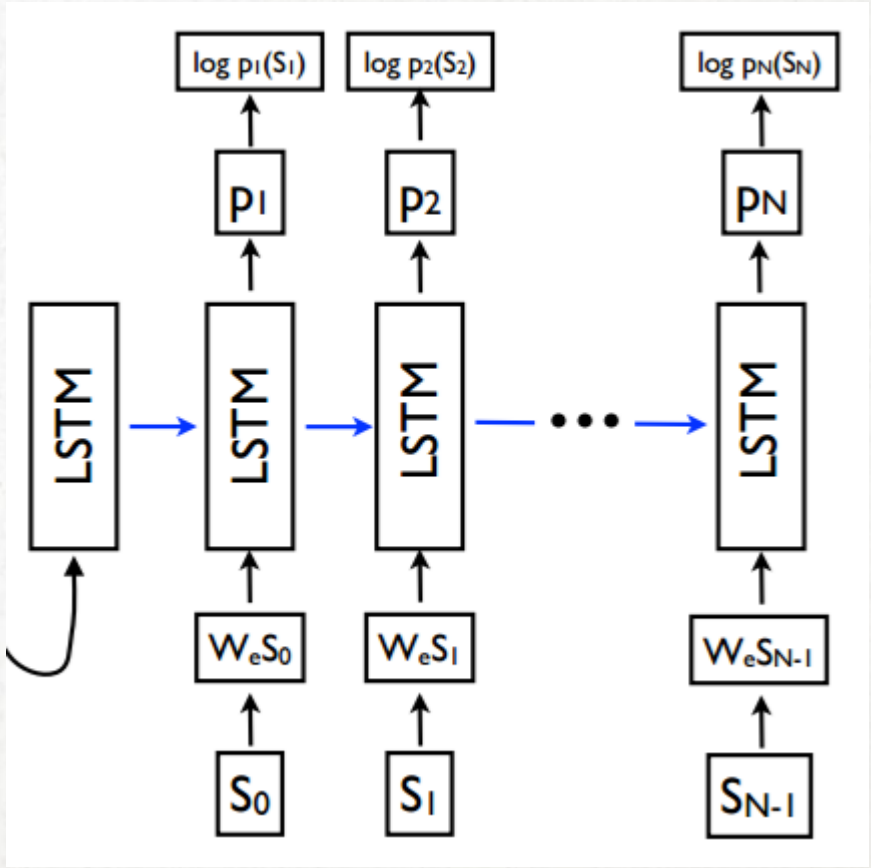


Image from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

WORD RNN

Two-layer LSTM with hidden size H = 512





Figures from O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.

EVALUATION AND EXPERIMENT

Dataset comprised of 19,551 image and annotation pairs

- Images are from MS COCO and Visual Genome
- Annotation were collected on Amazon Mechanical Turk
- Broken down to 14,575 training, 2,487 validation, and 2,489 testing images

Baselines:

- Sentence-Concat - Concatenates 5 sentence captions from a model trained on MS COCO captions
  - Purpose is to demonstrate difference between sentence-level and paragraph captions.
- Image-Flat – NeuralTalk
- Template – similar to BabyTalk
- Regions-Flat-Scratch – uses flat language model that's initialized from scratch
- Regions-Flat-Pretrained – same as above except using a pretrained language model

Model checkpoints are selected based on best combined METEOR and CIDEr score on validation set

QUANTITATIVE RESULTS

	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Sentence-Concat	12.05	6.82	31.11	15.10	7.56	3.98
Template	14.31	12.15	37.47	21.02	12.30	7.38
Image-Flat ([11])	12.82	11.06	34.04	19.95	12.20	7.71
Regions-Flat-Scratch	13.54	11.14	37.30	21.70	13.07	8.07
Regions-Flat-Pretrained	14.23	12.13	38.32	22.90	14.17	**8.85**
Regions-Hierarchical (ours)	**15.95**	**13.52**	**41.90**	**24.11**	**14.23**	8.69
Human	19.22	28.55	42.88	25.68	15.55	9.66

- Poor performance by Sentence-Concat shows the fundamental difference between single-sentence captioning and paragraph generation
- Template performed well on METEOR and CIDEr, but not so on BLEU-3 and BLEU-4. It indicates the template method is not good enough at describing relationships among objects in different regions
- Image-Flat and Regions-Flat-Scratch each improved the results further.
- Regions-Flat-Pretrained outperformed on all metrics, pre-training works
- The paper's method scored highest on all metrics except BLEU-4. Possibly due to Regions-Flat-Pretrained's non-hierarchical structure is better at exactly reproducing words immediately at the end and beginning of sentences

QUALITATIVE RESULTS

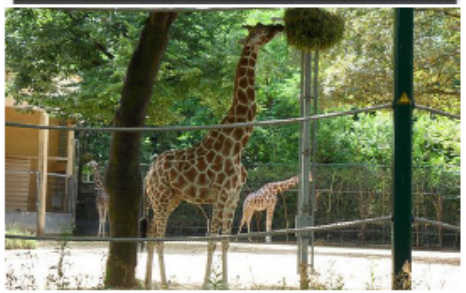
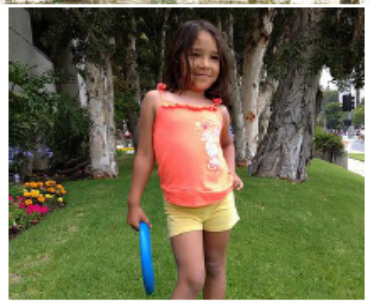
	Sentence-Concat	Template	Regions-Hierarchical
	A red double decker bus parked in a field. A double decker bus that is parked at the side of two and a road. A blue bus in the middle of a grand house. A new camera including a pinstripe boys and red white and blue outside. A large blue double decker bus with a front of a picture with its passengers in the.	There is a yellow and white bus, and a front wheel of a bus. There is a clear and blue sky, and a front wheel of a bus. There is a bus, and windows. There is a number on a train, and a white and red sign. There is a tire of a truck.	There are two buses driving in the road. There is a yellow bus on the road with white lines painted on it. It is stopped at the bus stop and a person is passing by it. In front of the bus there is a black and white bus.
	A man riding a horse drawn carriage down a street. Post with two men ride on the back of a wagon with large elephants. A man is on top of a horse in a wooden track. A person sitting on a bench with two horses in a street. The horse sits on a garage while he looks like he is traveling in.	People are riding a horse, and a man in a white shirt is sitting on a bench. People are sitting on a bench, and there is a wheel of a bicycle. There is a building with windows, and an blue umbrella. There are parked wheels, and a wheel. There is a brick.	A man is riding a carriage on a street. Two people are sitting on top of the horses. The carriage is made of wood. The carriage is black. The carriage has a white stripe down the side. The building in the background is a tan color.
	Two giraffes standing in a fenced in area. A big giraffe is is reading a tree. A giraffe sniffing the ground with its head. A couple of giraffe standing next to each other. Two giraffes are shown behind a fence and a fence.	Giraffes are standing in a field, and there is a standing giraffe. Tall green trees behind a fence are behind a fence, and there is a neck of a giraffe. There is a green grass, and a giraffe. There is a trunk of a tree, and a brown fence. there is a tree trunk, and white letters.	A giraffe is standing next to a tree. There is a pole with some green leaves on it to the right. There is a white and black brick building behind the fence. there are a bunch of trees and bushes as well.
	A young girl is playing with a frisbee. Man on a field with an orange frisbee. A woman holds a frisbee on a bench on a sunny day. A young girl is holding a green green frisbee. A girl throwing a frisbee in a park.	A girl is holding a tennis racket, and there is a green and brown grass. There is a pink shirt on a woman, and the background. The woman with a hair is wearing blue shorts, and there are red flowers. There are trees, and a blue frisbee in an air.	A woman in a red shirt and a black short short sleeve red shorts is holding a yellow frisbee. She is wearing a green shirt and white pants. She is wearing a pink shirt and short sleeve skirt. In her hand she is holding a white frisbee and a hand can be seen through it. Behind her are two white chairs. In the background is a large green and white building.

Figure 3. Example paragraph generation results for our model (Regions-Hierarchical) and the Sentence-Concat and Template baselines. The first three rows are positive results and the last row is a failure case.

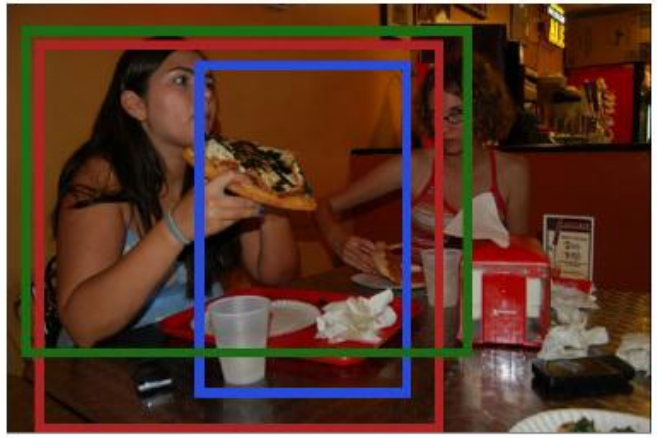
PARAGRAPH LANGUAGE ANALYSIS

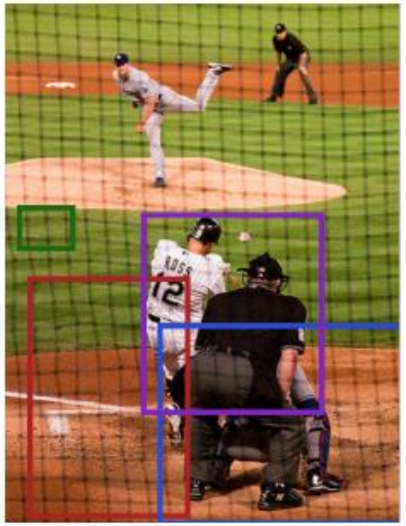
	Average Length	Std. Dev. Length	Diversity	Nouns	Verbs	Pronouns	Vocab Size
Sentence-Concat	56.18	4.74	34.23	32.53	9.74	0.95	2993
Template	60.81	7.01	45.42	23.23	11.83	0.00	422
Regions-Hierarchical	70.47	17.67	40.95	24.77	13.53	2.13	1989
Human	67.51	25.95	69.92	25.91	14.57	2.42	4137

- Similar average length and variance as human descriptions. The other 2 models fell short especially on variance of length, i.e. robotic
- Paper's method used more verbs and pronouns than the other automatic methods, and performed close to humans. That shows the robustness of describing actions and relationships in an image, and keep track of context among sentences
- Lots of room for improvement on Diversity for automatic methods

EXPLORATORY EXPERIMENT

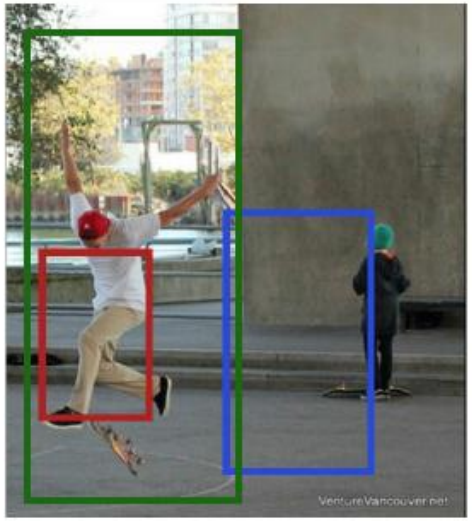
A young girl is sitting at a table in a restaurant. She is holding a hot dog on a bun in her hands. The girl is wearing a pink shirt and has short hair. A little girl is sitting on a table.





This is an image of a baseball game. The batter is wearing a white uniform with black lettering and a red helmet. The batter is wearing a white uniform with black lettering and a red helmet. The catcher is wearing a red helmet and red shirt and black pants. The catcher is wearing a red shirt and gray pants. The field is brown dirt and the grass is green.

Two men are standing on a skateboard on a ramp outside on a sunny day. One man is wearing black pants, a white shirt and black pants. The man on the skateboard is wearing jeans. The man's arms are stretched out in front of him. The man is wearing a white shirt and black pants. The other man is wearing a white shirt and black pants.



This is a sepia toned image on a cloudy day. There are a few white clouds in the sky. The tower has a clock on it with black numbers and numbers. The tower is white with black trim and black trim. the sky is blue with white clouds.

Figure 4. Examples of paragraph generation from only a few regions. Since only a small number of regions are used, this data is extremely out of sample for the model, but it is still able to focus on the regions of interest while ignoring the rest of the image.

THANK YOU!