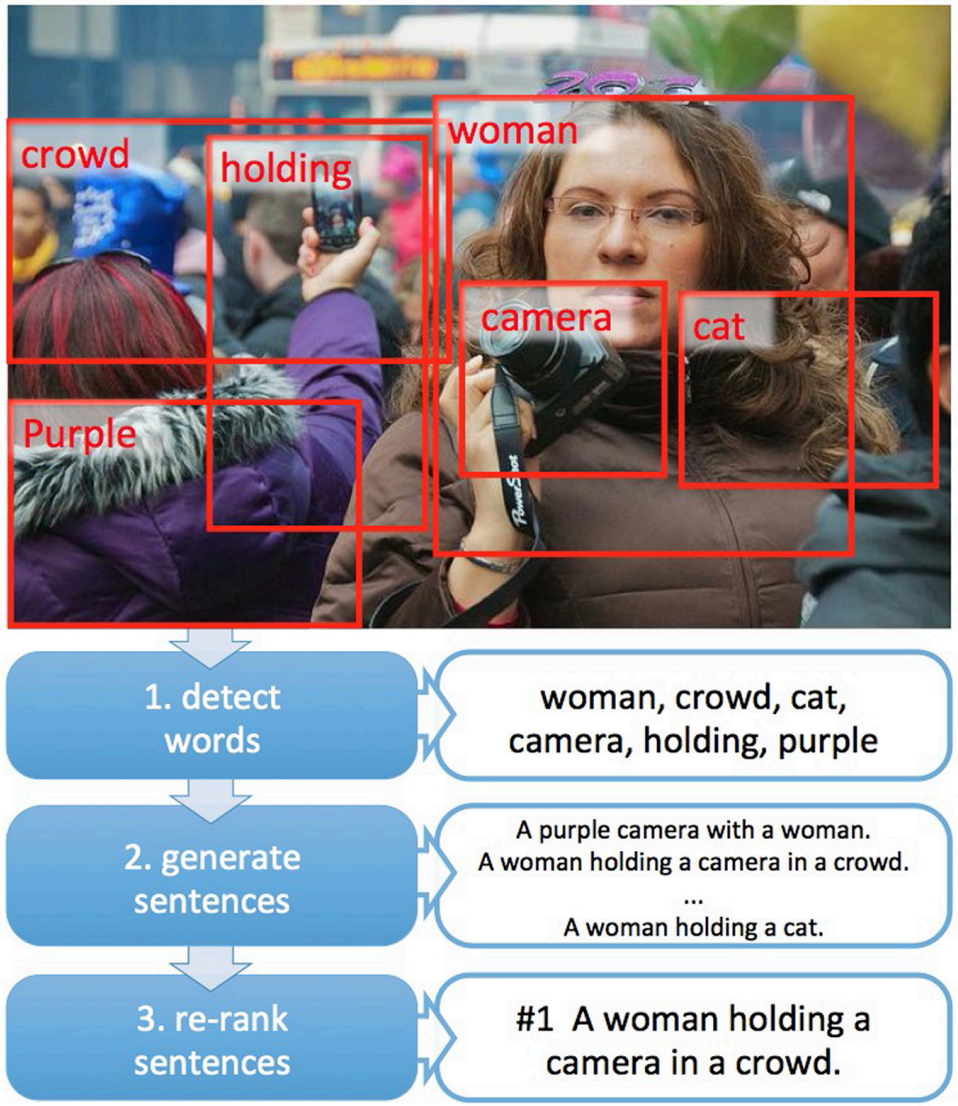
Descriptive Image Captioning

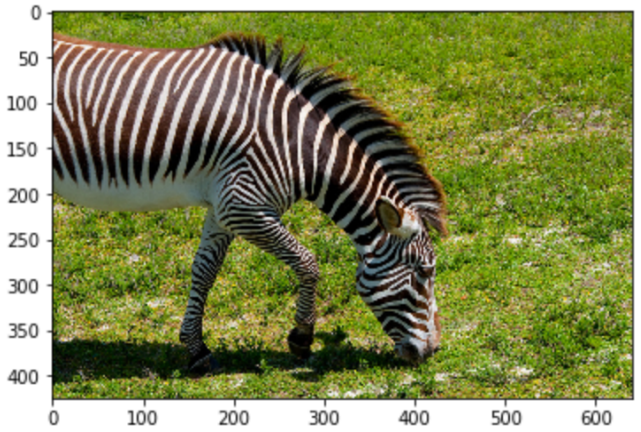
w266 NLP Final Project (Summer 2019)

Kenneth Chen
kl682@berkeley.edu

Image and caption

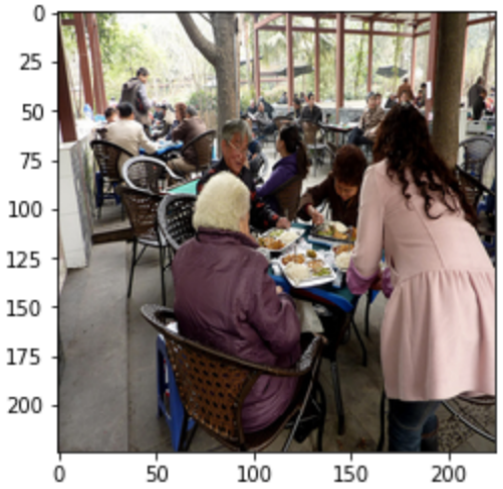


MS COCO2017 (Common Object in Context) dataset



1. A zebra grazing on lush green grass in a field.
2. Zebra reaching its head down to ground where grass is.
3. The zebra is eating grass in the sun.
4. A lone zebra grazing in some green grass.
5. a Zebra grazing on grass in a green open field.

Flickr dataset



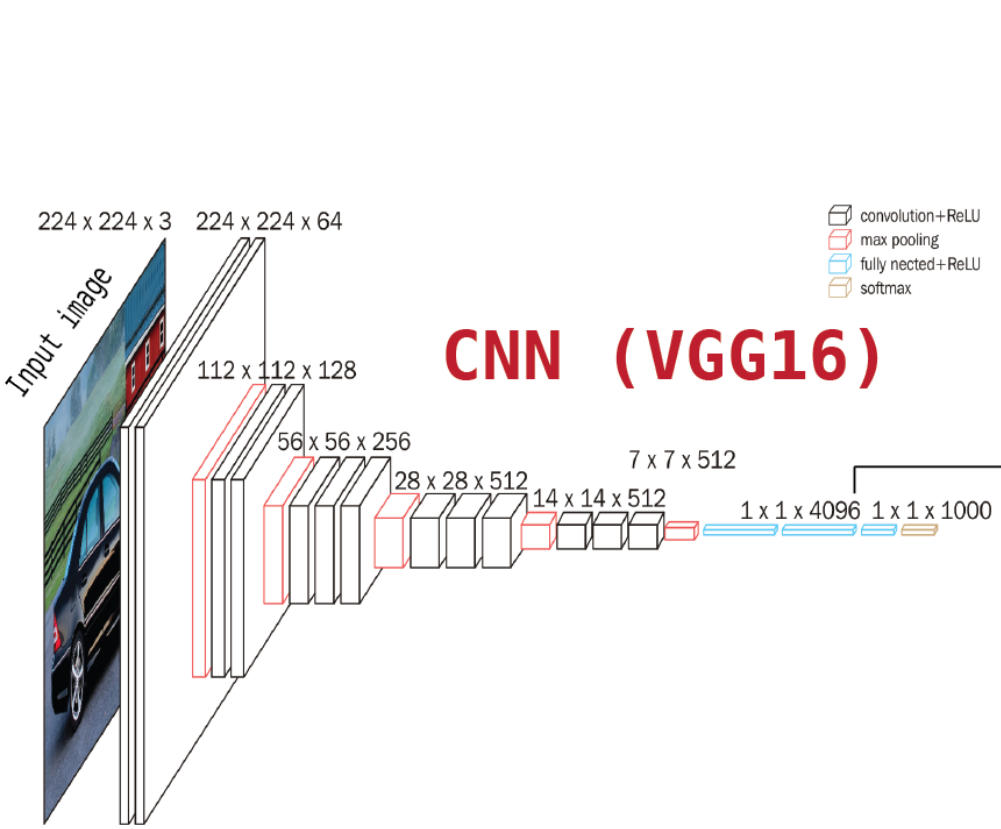
1. Even though it 's rather cool for outdoor dining , many people are enjoying the cafeteria-style food .
2. Several parties of people are seated in a dining room , eating food .
3. A woman in pink serves food two Asian people in an elderly home .
4. A group of people are having lunch in a crowded cafe .
5. Group having a meal in an outdoor setting .

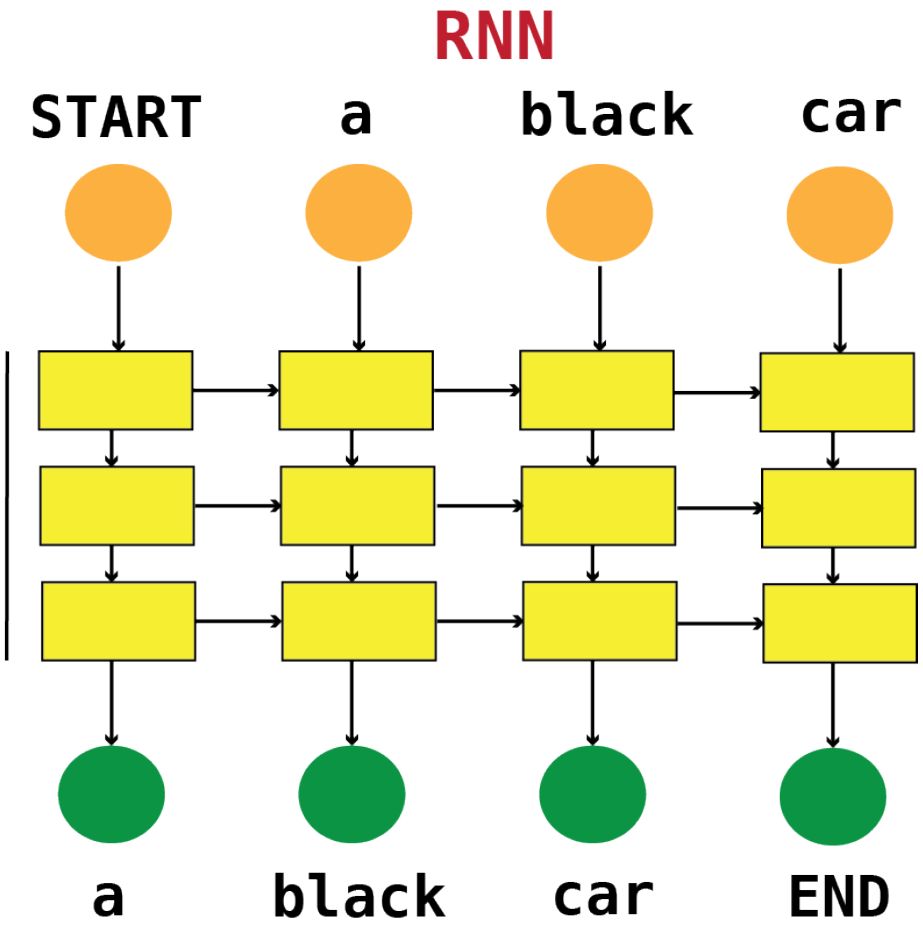
Multimodal Recurrent Neural Network (MRNN)

build model

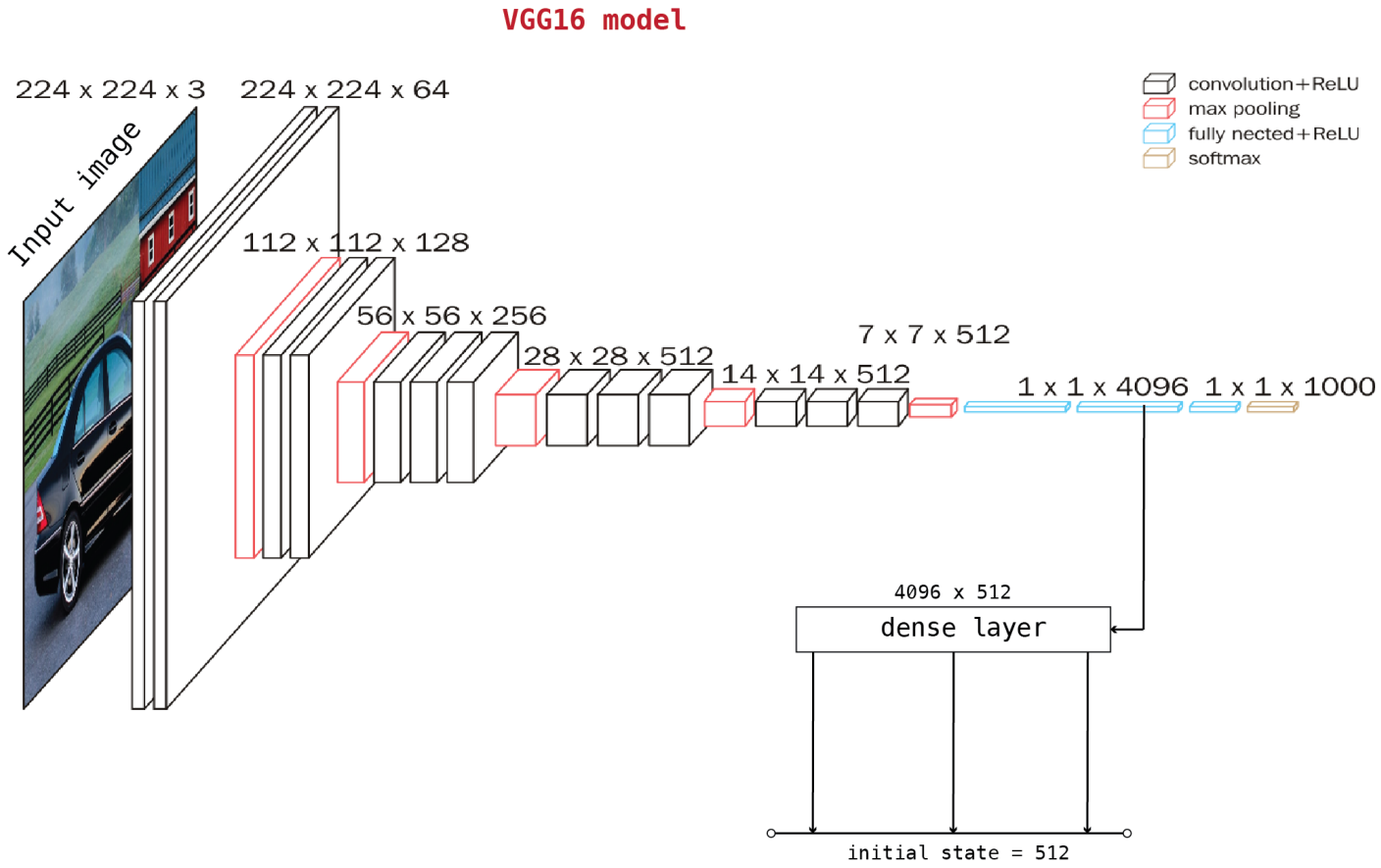
process images

process captions





Processing Images (COCO2017, Flickr8k, Flickr30k)



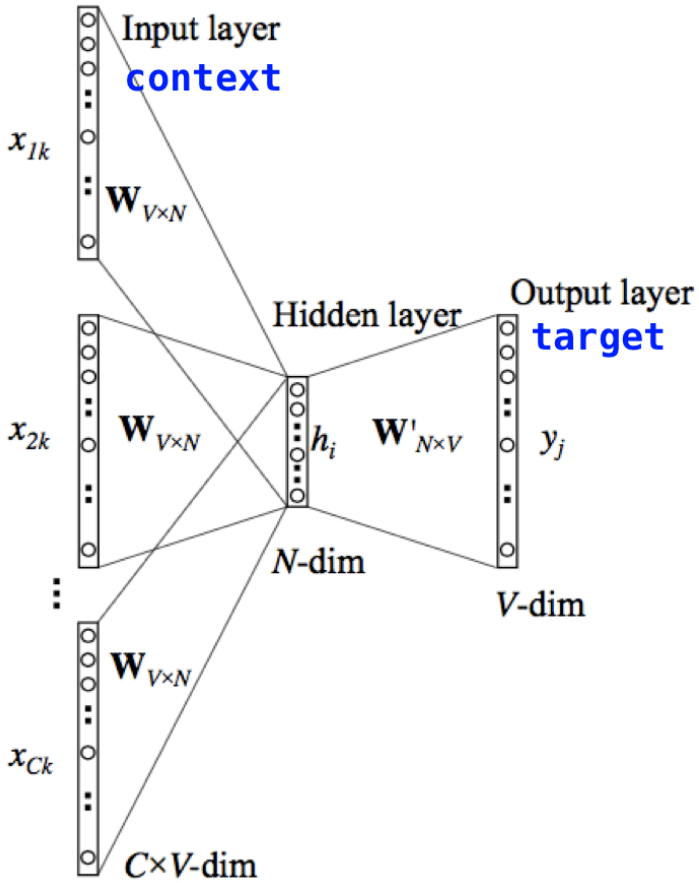
COCO2017 (Nvidia k80 2 hours)
train = 118,287 images
val = 5000 images

Flickr8k
train = 6000 images
val = 2000 images

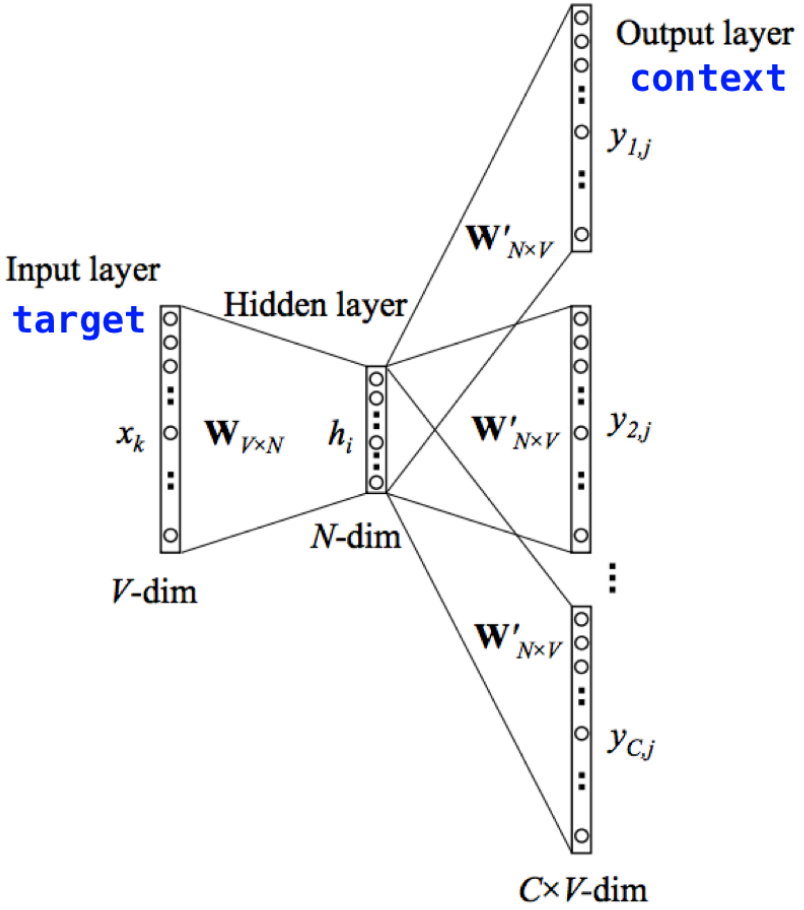
Flickr30k
train = 20000 images
val = 10000 images

Text processing

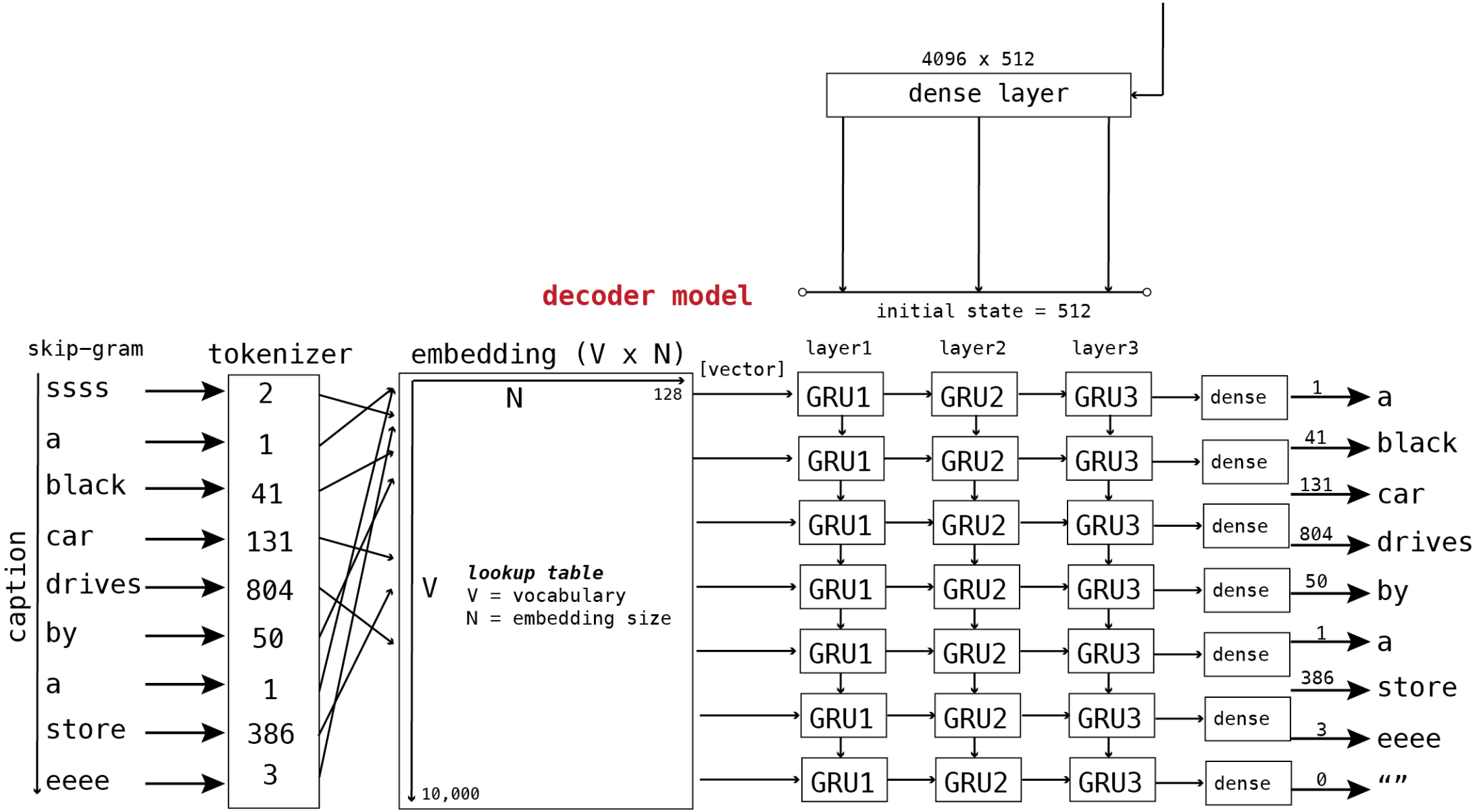
CBOW



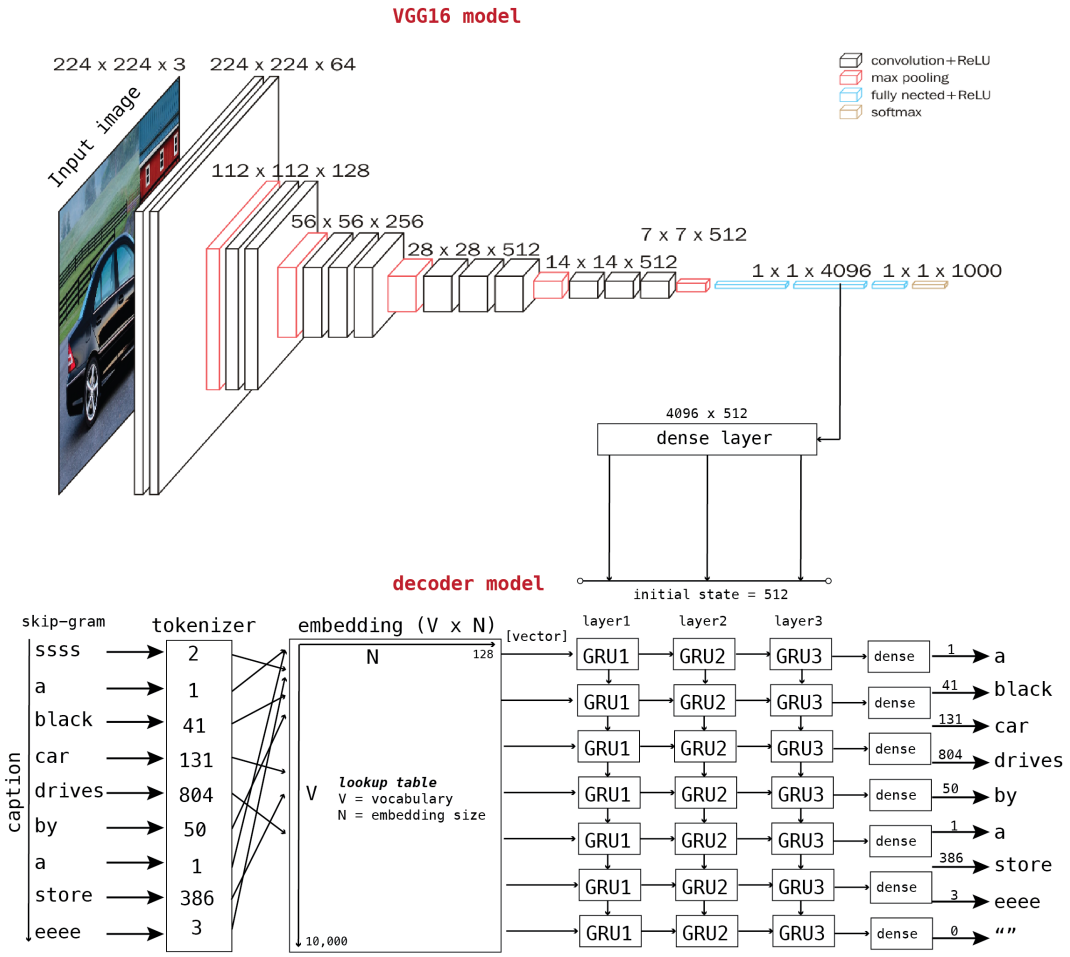
skip-gram



Process captions for RNN (decoder)



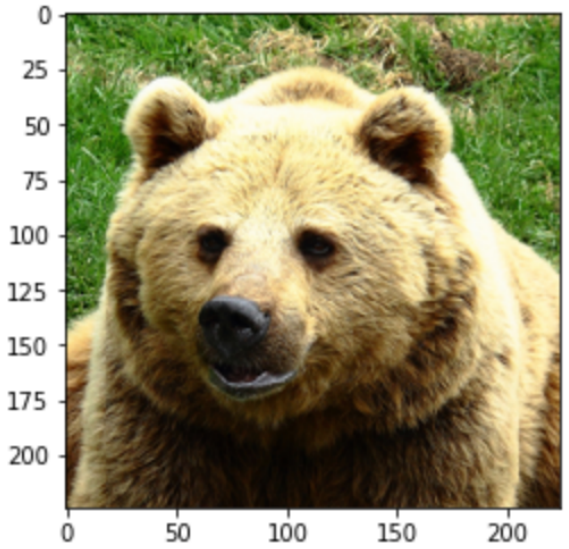
Build the MRNN model



Checkpoints

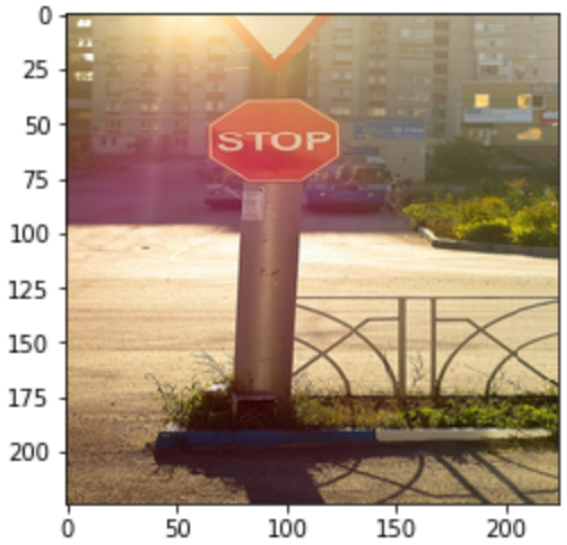
- checkpoint_w266_2caps_20epoch.keras
- checkpoint_w266_2caps_40epoch.keras
- checkpoint_w266_5GRU.keras
- checkpoint_w266_flickr8k_1GRU_40epoch.keras
- checkpoint_w266_flickr30k_1GRU_50epoch.keras
- checkpoint_w266_flickr30k_1GRU_70epoch.keras
- checkpoint_w266_flickr30k_2GRU_80epoch.keras
- checkpoint_w266_flickr30k_3GRU_200epoch.keras
- checkpoint_w266_flickr30k.keras
- checkpoint_w266.keras

https://drive.google.com/drive/folders/1f9_EgjJH6hUhvkZreywBhE0-p7i0ZdfA?usp=sharing



Predicted caption:
5 epochs = a a aa a
10 epochs = a brown bear sitting
20 epochs = a brown bear is sitting in a grassy field

True captions:
1. A big burly grizzly bear is show with grass in the background.
2. The large brown bear has a black nose.
3. Closeup of a brown bear sitting in a grassy area.
4. A large bear that is sitting on grass.
5. A close up picture of a brown bear's face.



Predicted caption:
a stop sign at the end of a street

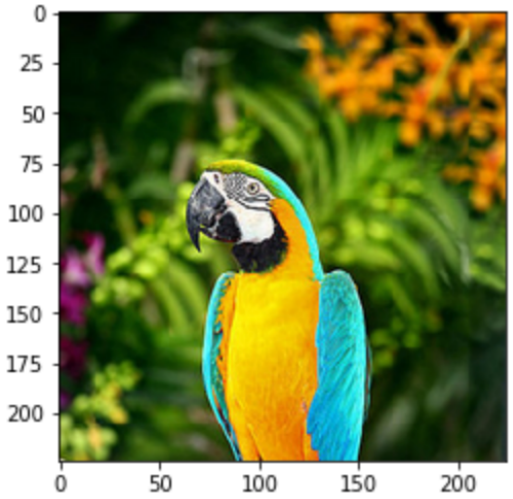
True captions:
A stop sigh on a post with two buses in the background.
an image of a stop sign and yield sign
A stop sign on post with a building in the background.
A stop sign sitting on top of a metal column.
a street sign with buildings in the background

MRNN trained on	COCO2017 [118287]			
test dataset	B-1	B-2	B-3	B-4
COCO[5000]	50.30	28.00	23.39	19.57

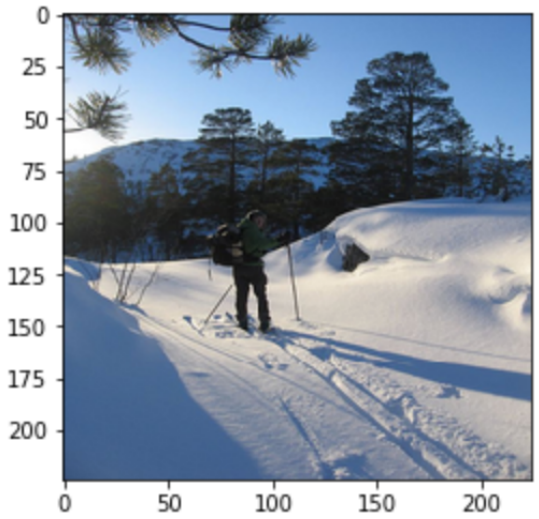
MRNN trained on	COCO2017 [118287]			
test dataset	cB-1	cB-2	cB-3	cB-4
COCO[5000]	50.30	36.32	31.02	27.18

MRNN trained on	COCO2017 [118287]					
test dataset	R-1	R-2	R-3	R-4	R-L1	R-W
COCO[5000]	30.88	8.25	2.67	1.04	34.44	15.74

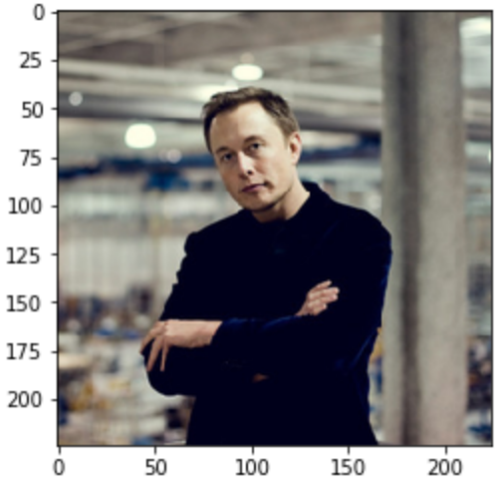
Predicted captions on non-COCO images using COCO trained model



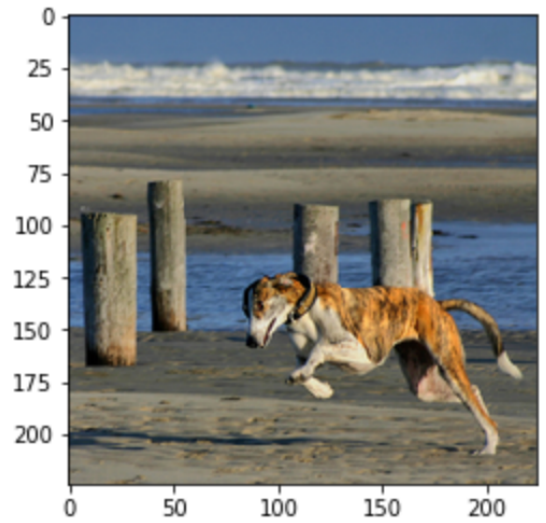
a small bird on a tree branch with a blurry background



a man is skiing down a snowy hill

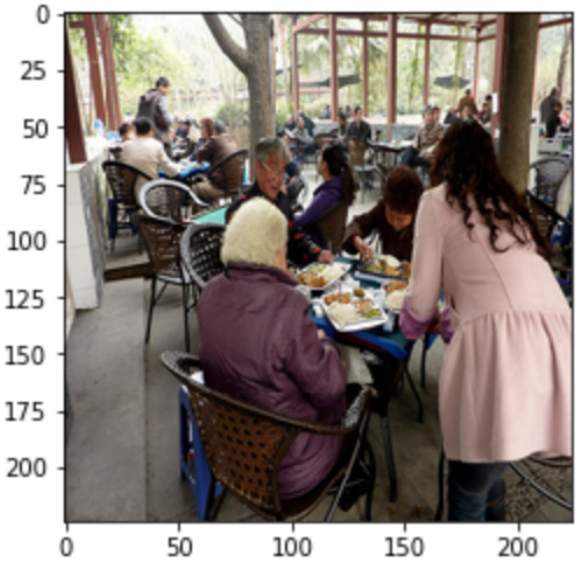


a man in a suit and tie standing in a boat



a man and a woman are playing frisbee on the beach

Flickr model



Predicted caption:
a man in a blue shirt is sitting on a stool

True captions:
1. Even though it 's rather cool for outdoor dining , many people are enjoying the cafeteria-style food .
2. Several parties of people are seated in a dining room , eating food .
3. A woman in pink serves food two Asian people in an elderly home .
4. A group of people are having lunch in a crowded cafe .
5. Group having a meal in an outdoor setting .

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

two young girls are playing with lego toy.

boy is doing backflip on wakeboard.

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

Table 2. Evaluation of full image predictions on 1,000 test images. **B-n** is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.

Karpathy and Fei Fei, 2015

Evaluation Metrics (BLEU and ROUGE scores)

MRNN trained on	COCO2017 [118287]				Flickr8k [6000]				Flickr30k [20000]			
test dataset	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
COCO[5000]	50.30	28.00	23.39	19.57	37.78	21.72	21.86	18.34	36.83	20.55	22.72	19.14
Flickr8k [8091]	40.40	23.20	21.83	18.39	44.78	23.41	21.44	17.89	45.54	23.65	22.17	18.80
Flickr30k [31783]	42.79	22.86	21.43	18.24	44.54	22.62	20.78	17.90	45.36	21.33	20.70	18.48

MRNN trained on	COCO2017 [118287]				Flickr8k [6000]				Flickr30k [20000]			
test dataset	cB-1	cB-2	cB-3	cB-4	cB-1	cB-2	cB-3	cB-4	cB-1	cB-2	cB-3	cB-4
COCO[5000]	50.30	36.32	31.02	27.18	37.78	27.31	25.18	22.85	36.83	25.97	24.40	22.49
Flickr8k [8091]	40.40	29.19	26.34	23.72	44.78	31.19	27.19	24.04	45.54	31.58	27.52	24.47
Flickr30k [31783]	42.79	29.93	26.58	23.83	44.54	30.56	26.46	23.55	45.36	29.88	25.74	23.11

MRNN trained on	COCO2017 [118287]						Flickr30k [20000]					
test dataset	R-1	R-2	R-3	R-4	R-L1	R-W	R-1	R-2	R-3	R-4	R-L1	R-W
COCO[5000]	30.88	8.25	2.67	1.04	34.44	15.74	27.37	5.19	1.12	0.37	31.22	13.92
Flickr8k [8091]	24.48	4.76	1.08	0.31	28.75	12.60	30.22	7.99	2.47	0.96	34.02	15.31
Flickr30k [31783]	23.64	5.01	1.29	0.42	27.77	11.74	27.34	6.95	2.14	0.87	30.91	12.96

Summary

- Abstract captionining makes it harder for model to achieve 100% evaluation scores
- Region Proposal Network (RPN) could be able to retrieve more descriptions in Flickr dataset
- 3 RNN layers improve the model better than 1 layer
- Embedding layer with random initialization is better than word2Vec (Karpathy & Fei Fei, 2015)

Implemention of Image captioning

- Surveillance camera
- accidents alert, activity involved