Project Report 2: Use Show And Tell Model to Generate Caption

Yue
yzk4v@mail.umkc.edu
University of Missouri-Kansas city
School of Computing & Engineering
Kansas City, MO 64110, USA

Abstract

In recent years, the development of computer power and the big data application make the machine learning and deep learning grow fast. Convolutional neural network is usually used for visual recognition, recurrent neural network is used for human language analysis some times. The show and tell model combine the two neural network for generating captions for images. We train the model use the data set of images and corresponding captions that produced by human being. First it use CNN to classify the features in the image, and encode it into vectors. Then it uses RNN to generate the captions from the vector of last step, it decodes the vector by using the model that learn from the caption then generate it word by word.

In this report, first introduce several approach that the model use, such as nature language processing (NLP) for human language analysis, convolutional neural network(CNN) for image classification and encode, recurrent neural network(RNN) for caption generation and decode, Scale Invariant Feature Transform(SIFT) for feature generation. And some explanation of supervised and unsupervised learning.

Supervised learning is to train the model by learning from the input that is be classified, and output the category by given a test input.Unsupervised learning is different that the data is not labeled. It is used for classification and decrease or increase the cost function.

Then we will show the result that produced by these model and approaches. First the proposed work of the network respectively, then implement them by combining these model, finally we evaluate the result and compare with other related work.

Introduction

Since the theme of project is animals, we don’t need to train the whole dataset for the

model, so we select several images and the corresponding captions that match the theme to train the model.

First we prepare the dataset for the model, including image and corresponding captions, then train the model with the pre-trained checkpoint, finally generate caption for the test data.

Related work

There are many approaches and model in deep learning that affect the accuracy of the result. Some neural network such as CNN[1], RNN[2], RCNN[3]. And there are many usage of these deep learning network like NLP[4], show and tell[5]. There are also some improvement like SIFT[6] for CNN, and LSTM[7] for RNN.

[8] incorporate two model to equip classification of low resolution pictures by combining convolutional high resolution and convolutional grained analysis.

[9] fix the model to resolve the problem of domain shift and learn the regression prototype by a zero-shot method.

[10] transfer unseen category to implanted space of seen category by fake labels without data loss.

Proposed Work

NLP[4]

Advances in natural language processing

In early days, the research about language is more about analyzing the linguistic structure, the research of today is more about applications of real-world, such as translation and discerning sentiment.

Computational linguistics grows a lot because of

1) Computing power increases a lot
2) A lot more linguistics data
3) The development of machine learning
4) The progress of the language understanding

In the before, people wrote the rules of human language for the computer, but it is a hard task. After 1990s, because of the development of big data and machine learning, we can train and build computational models with empirical language data.

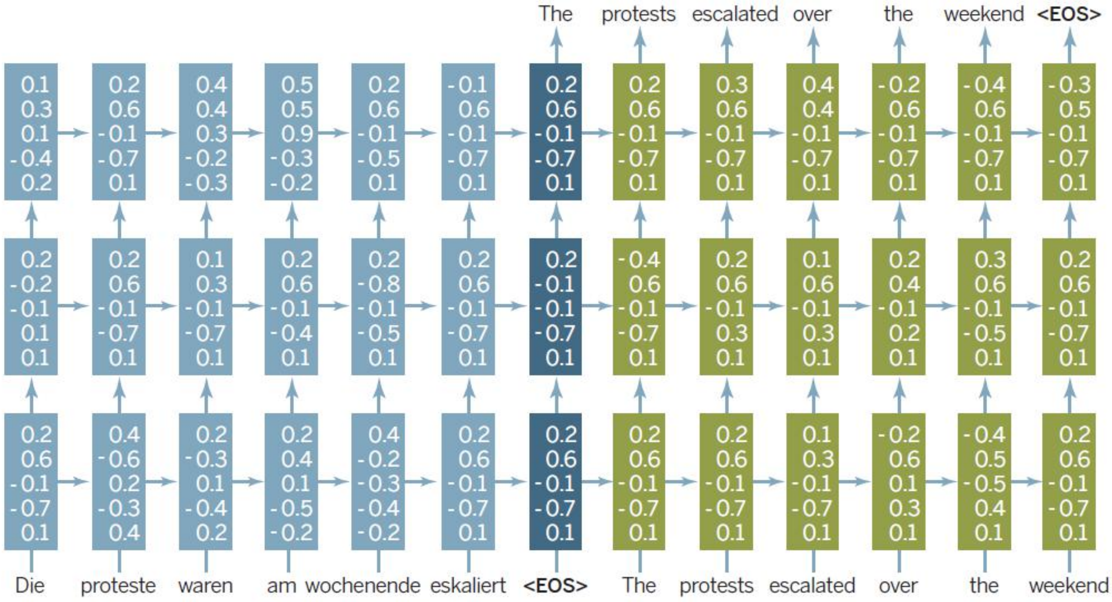
But there is a limitation that it only works on some high-resource languages like English, some low-resource languages like Punjabi, is spoken by millions people but NLP is not available for them.

Machine translation not only need to generate correspond sentence, but also need to understand the context for ambiguities.

The arrow shows the computation matrix multiplication unit with nonlinear transformation.

First it encodes the input(left blue bar), this contains a state of the portion input which is updated after every new word(horizontal arrows).

At the end of encoding(middle dark blue bar), it starts to generate the output translation from the state by using the model(right green bar). every word is fed by the input of each step.



An example of deep recurrent neural network

Feature Generation[6]

Scale Invariant Feature Transform

SIFT is used to discover useful pixel, the scale-invariance means the pixel is retained when scaling transforms, and it's transformed in concordance with the scaling. So this useful pixel is used to regularize the neighbour pixel with regard to scaling diversity. The pixel will also be constant when rotates.

In order to be scale constant, the scope of the neighbour pixel should be standardized within scale-invariant.

In order to be constant when rotates, the primary position around the neighbour is decided by the slope vectors around the neighbour. This is used for determining the grid where the diagram is computed with the primary position.

In order to be comparison constant, the SIFT is standardized to unit amount. So that the input of diagram can be invariable.

With the affine of picture severity near the pixel. This can develop the robustness of the model.

When SIFT is used in two diverse pictures, it can match the point reciprocally to find the pixel of other picture that make the Euclidean distance minimal.

SIFT matching is the most advanced method for recognition when combine the diverse item.

When SIFT if used for category or object classification, the analysis demonstrate that the SIFT over dense grids has better result than it over the sparse area. The reason is that the performance over the dense grid can generate more information than sparser area.

Although identify the category that already showed before can be done easily by SIFT, to classify an unseen category is still a difficult problem.

Show and tell[7]

The show and tell model can recognize the object in the image, and show the relationship between them, then describe it using nature language. It is an encoder-decoder NN model. First it encodes the image to a representation, then it decodes the representation to a caption.

In the encoding step, it uses CNN, CNN can embed the image to fixed-length vector, this vector will became the input of the decoding step.

In the decoding step, it uses RNN with LSTM to generate the representation to natural language caption.

Machine translation: input a sentence, make the probability that the translation is correct maximal.

So we can use the same approach, input a image, use CNN to generate the object, and use RNN to "translate" it into description by maximizing the probability of the correction of description.

Convolutional neural network[1]

We use CNN for image tasks instead of just increasing the hidden layers of normal neuron network, because of the limitation of computational power and the time for training, as well as the problem of overfitting.

Convolutional neural network is mostly used for image recognition and classification. Artificial Neural Networks is composed by associated neuron nodes, train the model by learning from the input dataset and optimize the output.

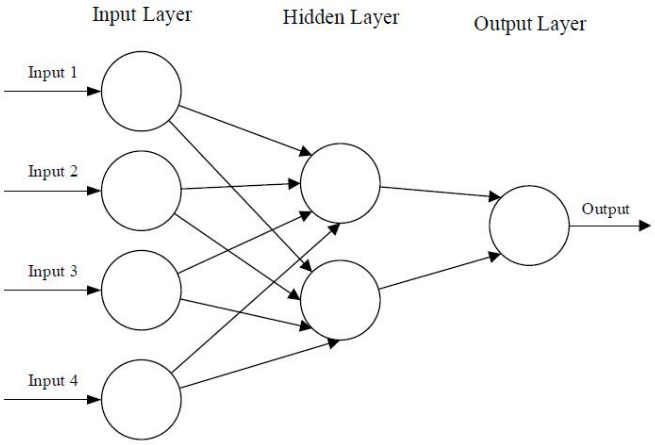
Normally the input is vector with many dimensions, it will be allocated to the hidden layers. Then the hidden layers will be determined by the preceding layers, and evaluate the progress and loss with the random change. This is called learning.

Supervised learning is to train the model by learning from the input that is be classified, and output the category by given a test input.

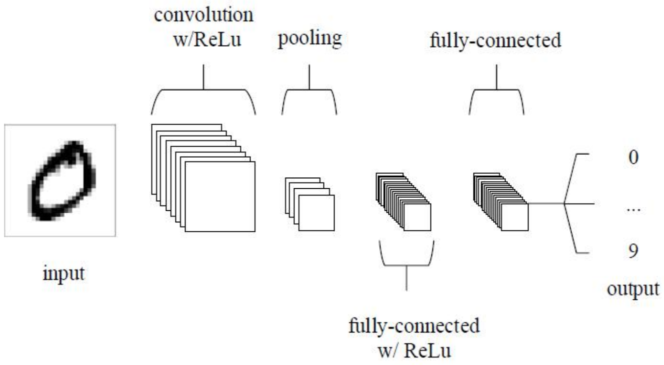
Unsupervised learning is different that the data is not labeled. It is used for classification and decrease or increase the cost function.

CNNs are mostly used for object recognition in images, it encode the features of the image to vector state, which can decrease the parameters of the model and make it more useful for image tasks.

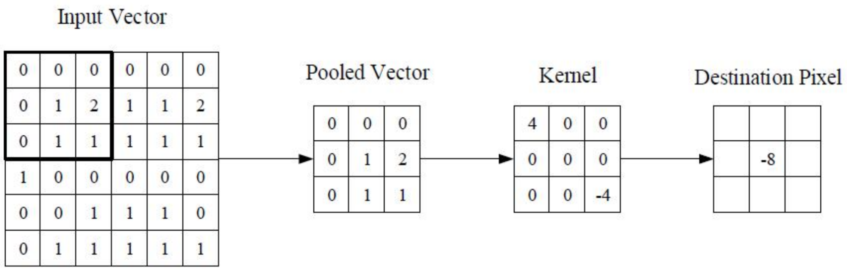
There are three dimensions of the input: height, width and depth. The output twill be like 1*1*n(n is the number of classes)



Neural network



Architecture of CNN



Input Vector

0	0	0	0	0	0
0	1	2	1	1	2
0	1	1	1	1	1
1	0	0	0	0	0
0	0	1	1	1	0
0	1	1	1	1	1

Pooled Vector

0	0	0
0	1	2
0	1	1

Kernel

4	0	0
0	0	0
0	0	-4

Destination Pixel

	-8	

Detail of the layer

CNN contains three kind of layers: convolutional layers, pooling layers and fully-connected layers.

1. The input layer includes the values of image pixel.

2. The convolutional layer decides the output that calculate the scalar within neighbour.

3. The pooling layer decrease the amount of parameters with activation function.

4. The fully-connected layer is the same as the neuron network that generate category scores from activation function.

Convolutional layer contains the learnable kernels. While the dimension of height and width may be small, but it will spread all over the input. The kernel will calculate with input vector and replace the weight by the result of itself and neighbour points.

Depth: The output capacity depth can be set by the amount of neurons with the layer to a equal input domain

Stride: Stride is length of the height and width dimension that we set to place approachable area.

Padding: zero-padding is to amplify the boarder and control the dimension of the output.

Parameter sharing: if a domain characteristic is useful at one area, then it can be used for another area

Pooling layer can diminish the dimension of the description which can decrease the amount of parameter and complexity of the model.

The neurons of fully-connected layer are straight associated with the neurons of the abutting layers but avoid to any other layers

Recurrent Convolutional Neural Network[3]

CNN is a structure based on feed, but recurrent associations are ample in visual system.

The input is unchanged, but the RCNN entity activities develop by the time.

The entity activity is adjusted by the nearby units.

Develop the RCNN by time will generate a random network with steady amount of parameters, just like other RNN.

CNN combine the back propagation algorithm to learn approachable simple pixel area

Fixed-point development is used for assumption, this is the connection between RCNN and sparse coding models.

Supervised learning approach can be combined to the unsupervised learning structure of sparse coding models.

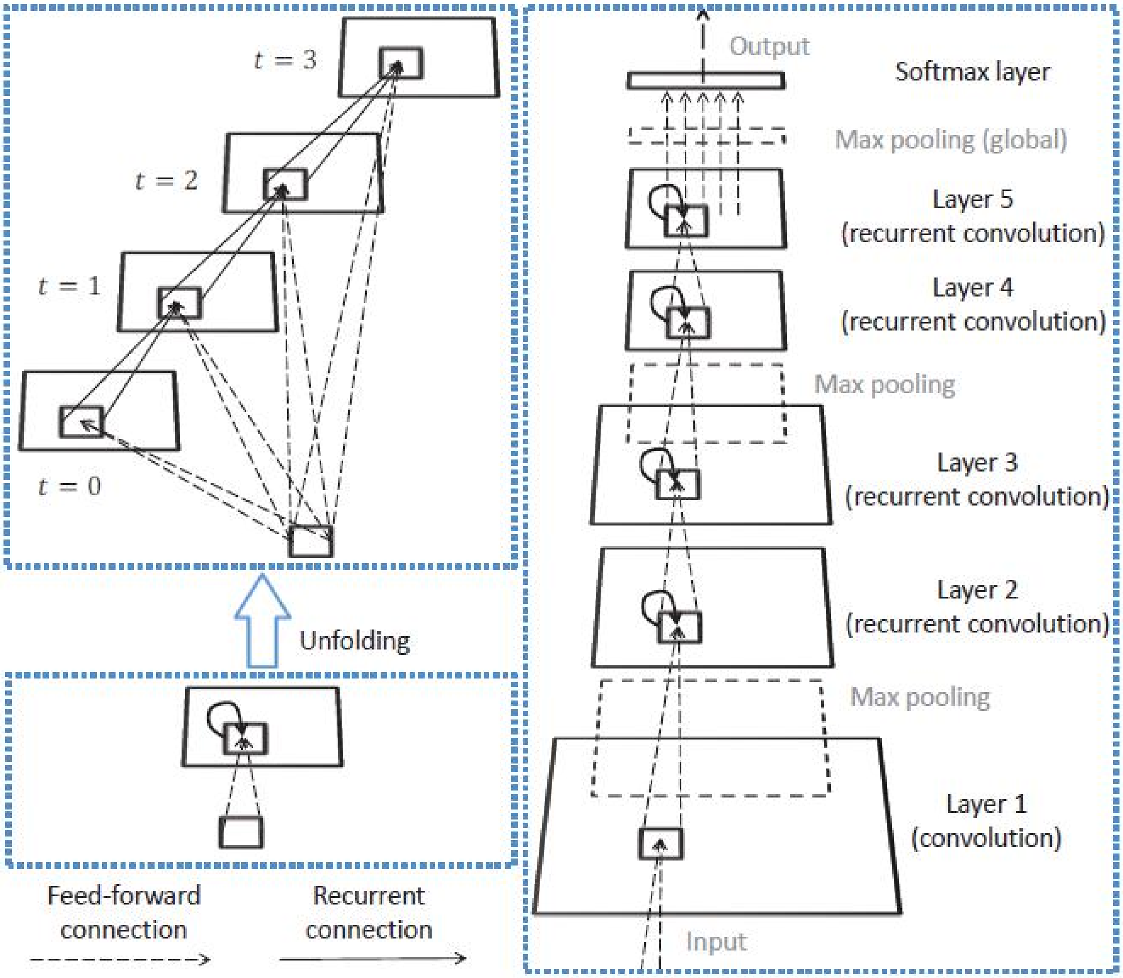
An important part of RCNN is recurrent convolutional layer. Develop the layer with time steps produce a T feed-onward subnetwork of T+1 extent.

There are various paths between input layer to output layer of subnetwork.

The longest one process all developed recurrent network(length = T+1)

The shortest one only process the feed-onward network

RCNN includes the a pile of recurrent convolutionals, with interleaved max-pooling layers by choice.



The comprehensive RCNN model.

Left: the recurrent convolutional layer is developed by time steps where T=3, resulting in a feed-onward subnetwork, its longest path is 4 and shortest path is 1. when t=0, only feed-onward occurs.

Right: the RCNN model includes one convolutional layer, 3 max pooling layers, a softmax layer and four recurrent convolutional layers.

For saving the space and time, first layer is regular feed-onward convolutional layer excluding recurrent networks, with max pooling layer next.

Only feed-onward network is between recurrent convolutional layer.

Use back propagation through time algorithm(BPTT algorithm) to train the model, it will minimize the intersect entropy loss operation. This is same as using regular back propagation algorithm at a time developed connection.

The advantages of RCNN:

It make the unit to combine the context in the large area at the current layer.

The recurrent network expand the depth but use the weight sharing to make the amount of changeable parameters unchanged.

The time developed RCNN is the CNN with various paths from input layer through output layer.

The longer path make the model learn very complex categories while training, while the shorter path can help gradient back propagation for training.

A few repetition of the changing action can generate outstanding result.

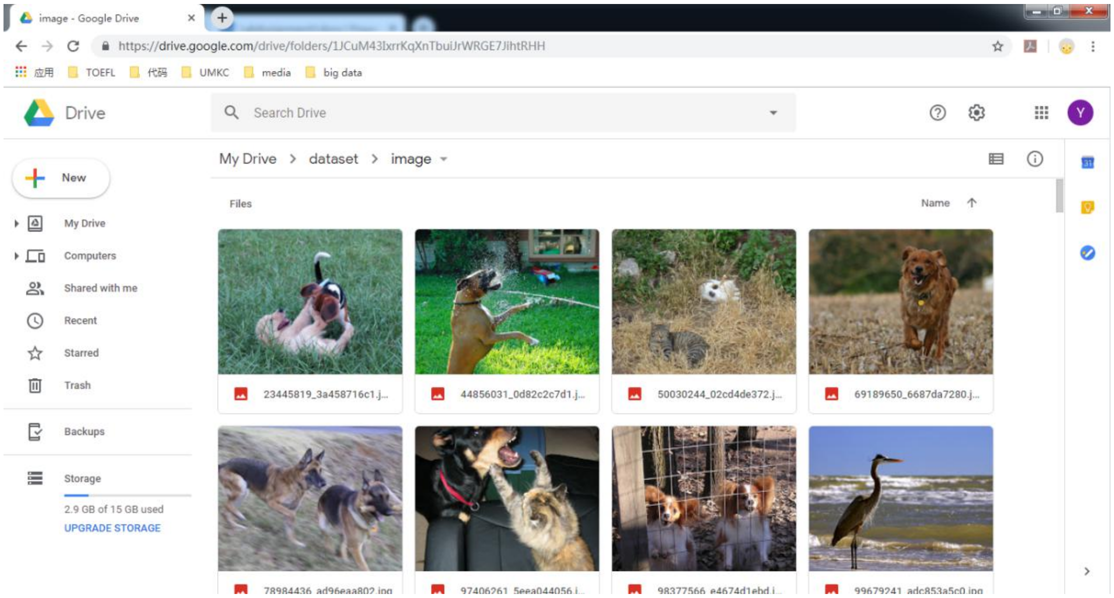
We can compare RCNN with other model to analyze the feature, such as the model that remove the recurrent network of RCNN which become a normal CNN, and remove the recurrent network only in the recurrent convolution layer

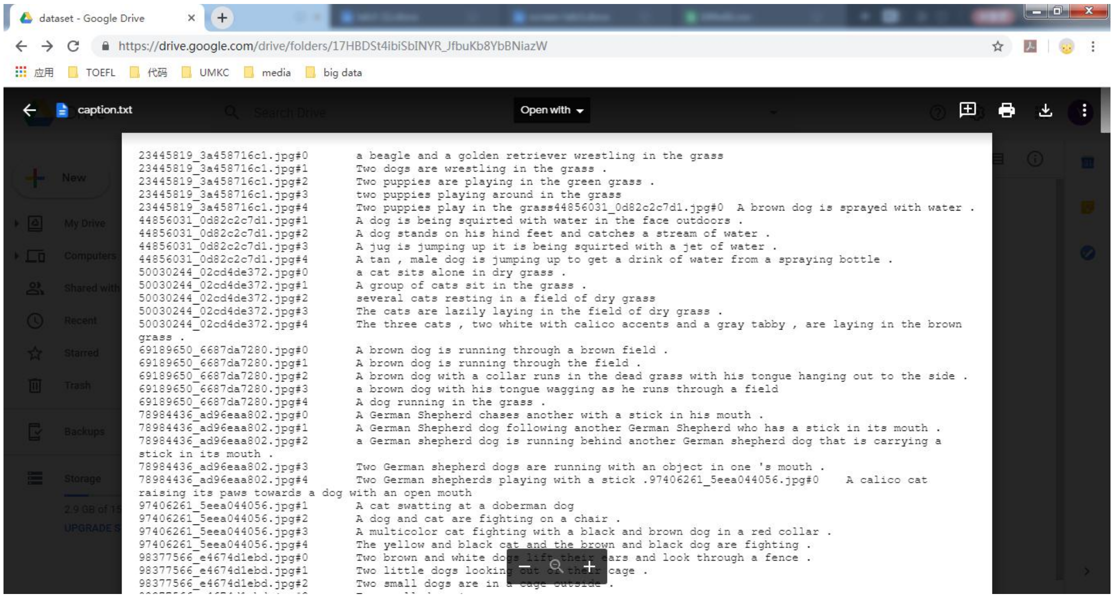
With the idea of synapses of human brain, people develop a neuron network to recognize object that called RCNN. It adds recurrent network to each convolutional layer in the feed-onward CNN.

The recurrent network expand the depth while still keep the amount of parameters constant by using the weight sharing

Implementation and Evaluation

1. Prepare the dataset for the model, including image and corresponding captions.

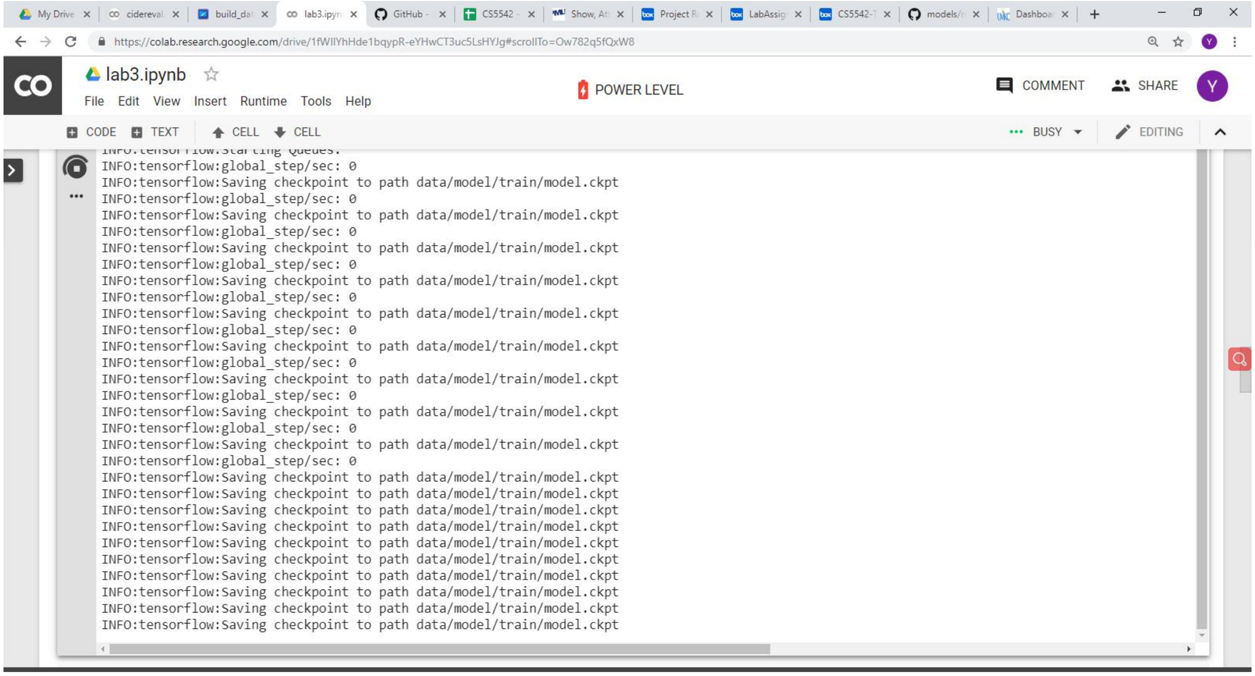




dataset - Google Drive
https://drive.google.com/drive/folders/17HBDSt4ibiSbINYR_JfbuKb8YbBNiazW
应用
TOEFL
代码
UMKC
media
big data
caption.txt
Open with

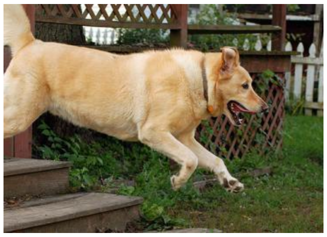
```
23445819_3a458716c1.jpg#0       a beagle and a golden retriever wrestling in the grass
23445819_3a458716c1.jpg#1       Two dogs are wrestling in the grass .
23445819_3a458716c1.jpg#2       Two puppies are playing in the green grass .
23445819_3a458716c1.jpg#3       two puppies playing around in the grass
23445819_3a458716c1.jpg#4       Two puppies play in the grass44856031_0d82c2c7d1.jpg#0  A brown dog is sprayed with water .
44856031_0d82c2c7d1.jpg#1       A dog is being squirted with water in the face outdoors .
44856031_0d82c2c7d1.jpg#2       A dog stands on his hind feet and catches a stream of water .
44856031_0d82c2c7d1.jpg#3       A jug is jumping up it is being squirted with a jet of water .
44856031_0d82c2c7d1.jpg#4       A tan , male dog is jumping up to get a drink of water from a spraying bottle .
50030244_02cd4de372.jpg#0       a cat sits alone in dry grass .
50030244_02cd4de372.jpg#1       A group of cats sit in the grass .
50030244_02cd4de372.jpg#2       several cats resting in a field of dry grass
50030244_02cd4de372.jpg#3       The cats are lazily laying in the field of dry grass .
50030244_02cd4de372.jpg#4       The three cats , two white with calico accents and a gray tabby , are laying in the brown
grass .
69189650_6687da7280.jpg#0       A brown dog is running through a brown field .
69189650_6687da7280.jpg#1       A brown dog is running through the field .
69189650_6687da7280.jpg#2       A brown dog with a collar runs in the dead grass with his tongue hanging out to the side .
69189650_6687da7280.jpg#3       a brown dog with his tongue wagging as he runs through a field
69189650_6687da7280.jpg#4       A dog running in the grass .
78984436_ad96eaa802.jpg#0       A German Shepherd chases another with a stick in his mouth .
78984436_ad96eaa802.jpg#1       A German Shepherd dog following another German Shepherd who has a stick in its mouth .
78984436_ad96eaa802.jpg#2       a German shepherd dog is running behind another German shepherd dog that is carrying a
stick in its mouth .
78984436_ad96eaa802.jpg#3       Two German shepherd dogs are running with an object in one 's mouth .
78984436_ad96eaa802.jpg#4       Two German shepherds playing with a stick .97406261_5eea044056.jpg#0    A calico cat
raising its paws towards a dog with an open mouth
97406261_5eea044056.jpg#1       A cat swatting at a doberman dog
97406261_5eea044056.jpg#2       A dog and cat are fighting on a chair .
97406261_5eea044056.jpg#3       A multicolor cat fighting with a black and brown dog in a red collar .
97406261_5eea044056.jpg#4       The yellow and black cat and the brown and black dog are fighting .
98377566_e4674d1ebd.jpg#0       Two brown and white dogs lift their ears and look through a fence .
98377566_e4674d1ebd.jpg#1       Two little dogs looking out of their cage .
98377566_e4674d1ebd.jpg#2       Two small dogs are in a cage outside .
```

2. Train the model with the pre-trained checkpoint.



```
INFO:tensorflow:Starting Queues.
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:global_step/sec: 0
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
INFO:tensorflow:Saving checkpoint to path data/model/train/model.ckpt
```

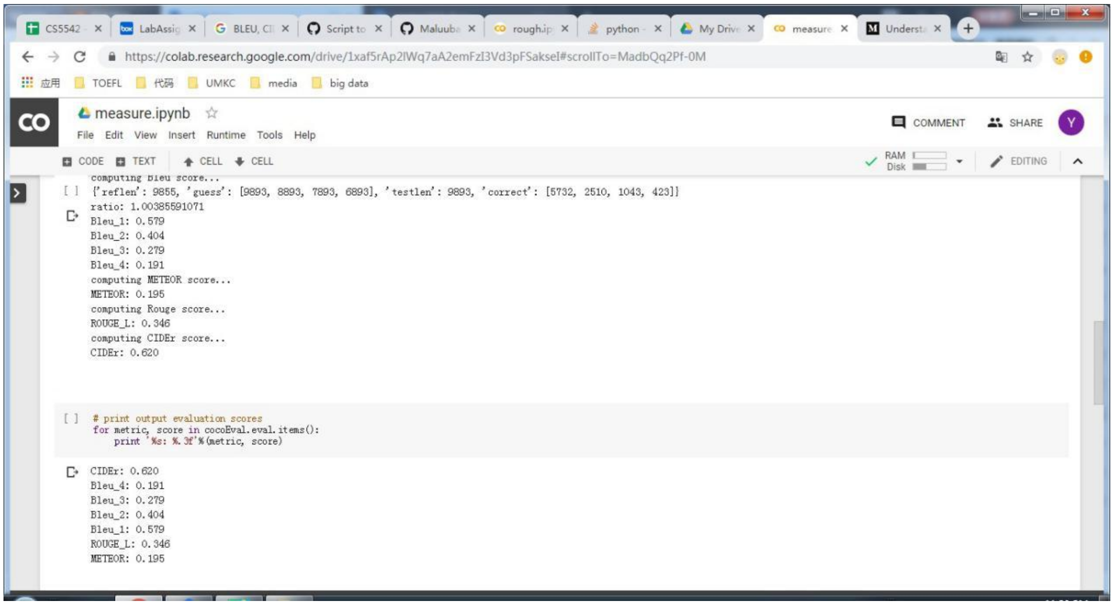
3. Generate caption for the test data



A blond dog runs down a flight of stairs to the backyard .

A dog jumps off the stairs .
A tan dog runs down a wooden staircase to the green grass .

4. accuracy in BLEU, CIDER, METEOR and ROGUE measures



```
computing Bleu score...
{'reflen': 9855, 'guess': [9893, 8893, 7893, 6893], 'testlen': 9893, 'correct': [5732, 2510, 1043, 423]}
ratio: 1.00385591071
Bleu_1: 0.579
Bleu_2: 0.404
Bleu_3: 0.279
Bleu_4: 0.191
computing METEOR score...
METEOR: 0.195
computing Rouge score...
ROUGE_L: 0.346
computing CIDEr score...
CIDEr: 0.620
```

```
# print output evaluation scores
for metric, score in cocoEval.eval.items():
    print '%s: %.3f'%(metric, score)
```

```
CIDEr: 0.620
Bleu_4: 0.191
Bleu_3: 0.279
Bleu_2: 0.404
Bleu_1: 0.579
ROUGE_L: 0.346
METEOR: 0.195
```

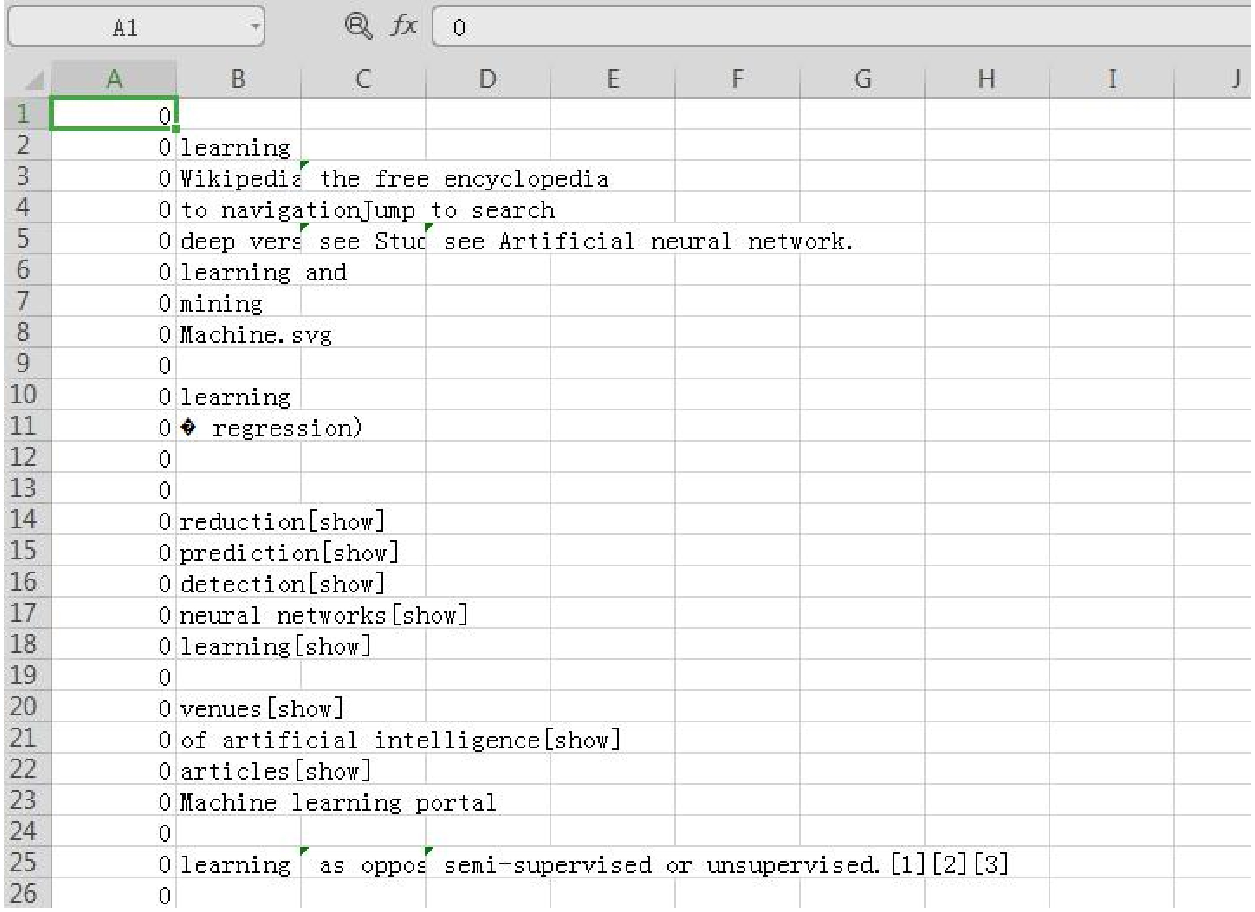
BLEU-4: 0.191
CIDER: 0.620
METEOR: 0.196
ROGUE: 0.346

```
#The im2txt use the MSCOCO dataset, we use our own dataset.
#set dataset path
tf.flags.DEFINE_string("image_dir", "data/image/",
                       "Image directory.")
tf.flags.DEFINE_string("captions_file", "data/caption.txt",
                       "Captions text file.")
```

II. Unsupervised learning by using clustering

```
//Use the kmeans library of spark
import org.apache.spark.mllib.clustering.KMeans
val kMeansModel=KMeans.train(tf,10,1000) //model
val WSSSE = kMeansModel.computeCost(tf)//Within Set Sum of Squared Errors
val clusters=kMeansModel.predict(tf) //use predict function of the model
```

Output



	A	B	C	D	E	F	G	H	I	J
1	0									
2	0	learning								
3	0	Wikipedia	the free encyclopedia							
4	0	to navigationJump to search								
5	0	deep vers	see Stud	see Artificial neural network.						
6	0	learning and								
7	0	mining								
8	0	Machine.svg								
9	0									
10	0	learning								
11	0	� regression)								
12	0									
13	0									
14	0	reduction[show]								
15	0	prediction[show]								
16	0	detection[show]								
17	0	neural networks[show]								
18	0	learning[show]								
19	0									
20	0	venues[show]								
21	0	of artificial intelligence[show]								
22	0	articles[show]								
23	0	Machine learning portal								
24	0									
25	0	learning	as oppos	semi-supervised or unsupervised.[1][2][3]						
26	0									

Conclusion

The model is neural image caption that use the combination of CNN and RNN to produce the corresponding the description of the given image. Given an image, this model will maximize the probability of the caption by training. We can see in the metrics result the model shows its accuracy. In the future, there will be some better metrics for evaluation to fit the goal.

Reference

[1]Keiron O’Shea and Ryan Nash, “An Introduction to Convolutional Neural Networks”, Dec 2015

[2] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, “SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS”, Mar 2013

[3] Ming Liang, Xiaolin Hu, ”Recurrent Convolutional Neural Network for Object Recognition”, 2015

[4] Julia Hirschberg, Christopher D. Manning, “Advances in natural language processing”, April 5, 2019

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, “Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge”, Sep 2016

[6] Tony Lindeberg, “Scale Invariant Feature Transform”, Stockholm, Sweden, 2015

[7] Haşim Sak, Andrew Senior, Françoise Beaufays, “Long Short-Term Memory

Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition", Feb 2014

[8] D. Cai , K. Chen , Y. Qian , J.-K. Kämäräinen , Convolutional low-resolution fine–grained classification, Pattern Recognit. Lett. 119 (2019) 116–171 .

[9] C. Luo , Z. Li , K. Huang , J. Feng , M. Wang , Zero-shot learning via attribute regression and class prototype rectification, IEEE Trans. Image Process. 27 (2) (2018) 637–648 .

[10] Y. Guo , G. Ding , J. Han , Y. Gao , Zero-shot learning with transferred samples, IEEE Trans. Image Process. 26 (7) (2017) 3277–3290 .