Dynamic Lexicon Generation for Natural Scene Images

Mentor : Sarthak Sharma
Bhavin Kotak - 2018201071
Ishan Tyagi - 2018201017
Jatin Paliwal - 2018202006
Vishal Bidawatka - 2018201004

LDA

Results

Objective

CNN

Thank You

Objective

- The problem is to generate contextualized lexicon given only visual information.
- For this, we exploit the correlation between visual and textual information in a dataset consisting of images and textual content associated with them.

Purpose

Approach

Purpose

- In most computer vision problems such custom lexicons are artificially created and provided to the algorithm as a form of predefined word queries. But, in real life scenarios lexicons need to be dynamically constructed.

Approach

Step:1-2

Step 3

Step 1: LDA

First, we learn a topic model using Latent Dirichlet Allocation (LDA) using as input a corpus textual information associated with scene images combined with scene text.

Step 2: CNN

STEP 2: We train a deep CNN model, based on the topic model, that is capable to produce on its output a probability distribution over the topics discovered by the LDA analysis directly from the image input.

Step 3 : Dictionary Re-ranking

- By the usage of the topic probabilities we can generate word rankings, i.e. a per-image ranked lexicon, for new (unseen) images.

LDA

- Topic model is a type of statistical model for discovering the abstract topics that occur in a collection of text documents.Here an image is a collection of different topics.
- Latent Dirichlet Allocation is a topic modeling technique that when given a text corpus, it defines the the relation of topics w.r.t. words and relation of document w.r.t. topics.

Example

Input Data

Model definitions

Example

Step 1 :

Document 0	Hi my name is Jatin
Document 1	Jatin Paliwal loves football
Document 2	Jatin is a student of IIIT Hyderabad

Step 2 :

	a	football	hi	hyderabad	iiit	is	jatin	loves	my	name	of	paliwal	student
Document 0	0	0	1	0	0	1	1	0	1	1	0	0	0
Document 1	0	1	0	0	0	0	1	1	0	0	0	1	0
Document 2	1	0	0	1	1	1	1	0	0	0	1	0	1
	a	**football**	**hi**	**hyderabad**	**iiit**	**is**	**jatin**	**loves**	**my**	**name**	**of**	**paliwal**	**student**

Step 3 :

	Topic 0	Topic 1	Topic 2
a	0.002	0.005	0.124
football	0.002	0.474	0.001
hi	0.165	0.005	0.001
hyderabad	0.165	0.005	0.001
iiit	0.002	0.005	0.124
is	0.328	0.005	0.001
jatin	0.002	0.005	0.370
loves	0.002	0.005	0.124
my	0.165	0.005	0.001
name	0.165	0.005	0.001
of	0.002	0.005	0.124
paliwal	0.002	0.474	0.001
student	0.002	0.005	0.124
	Topic 0	Topic 1	Topic 2

	Topic 0	Topic 1	Topic 2
Document 0	0.722	0.056	0.222
Document 1	0.067	0.467	0.467
Document 2	0.292	0.042	0.667
	Topic 0	Topic 1	Topic 2

Input data

- We used MS COCO data set, that has captions with respect to a image id in the following way.
- http://images.cocodataset.org/annotations/annotations_trainval2014.zip

```
['A bicycle replica with a clock as the front wheel.',
 'The bike has a clock as a tire.',
 'A black metal bicycle with a clock inside the front wheel.',
 'A bicycle figurine in which the front wheel is replaced with a clock',
 'A clock with the appearance of the wheel of a bicycle ']
```

An example of input image and its corpus to LDA

```
[(0,
  '0.072*"street" + 0.069*"clock" + 0.057*"build" + 0.043*"sign" + 0.028*"tower" + 0.028*"park" +
0.023*"large" + 0.021*"road" + 0.019*"light" + 0.018*"motorcycle" + 0.017*"stop" + 0.012*"tall" +
k" + 0.011*"near" + 0.011*"traffic" + 0.011*"white" + 0.010*"bike" + 0.010*"pole" + 0.009*"cars" -
+ 0.008*"brick" + 0.008*"walk" + 0.008*"outside" + 0.007*"blue" + 0.007*"green" + 0.006*"bicycle"
```

```
In [112]: lda_model[bow_corpus[0]]
Out[112]: [(0, 0.72454524), (1, 0.062070537), (4, 0.1828729)]
```

The Output of LDA for the previous image

Model used : Gensim

Important Parameters

- Number of Topics
- Alpha
- Beta

CNN

- Convolutional Neural Network (CNN) will be used to predict the probability of topic given image.
- This way our method is able to generate contextualized lexicons for new (unseen) images directly from their raw pixels, without the need of any associated textual content.

Procedure

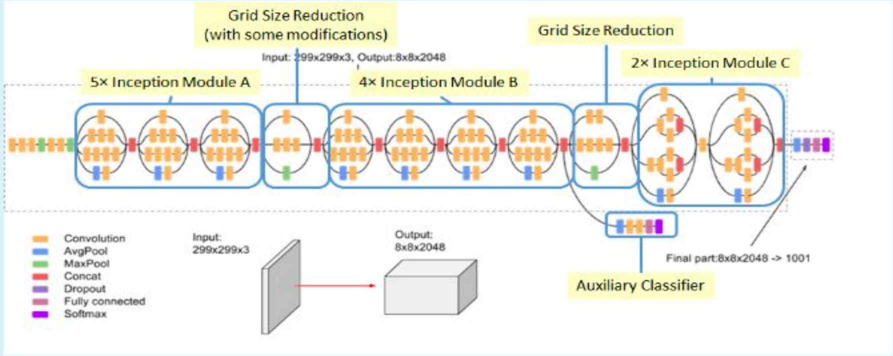
Procedure

Input Data

- The output of LDA , P(topic|text) i.e. Probability of topic given a text caption of an image is used as labels (y) and image as an input feature (x)
- As a result, on the basis of raw pixels only the network can generate topic probabilities.

Model Used : InceptionV3

- We used pre-trained convolutional neural network Inception V3 with keras module.
- As the part of transfer learning , we removed the last layer of the inception model and and trained the fully connected neural network which takes input from inception model.



Model Parameters

Inception model +
dense layer of 256 nodes with Relu activation +
dense layer of 128 nodes with Relu acitvation +
Output layer with nodes equal to number of topics and Softmax activaiton.

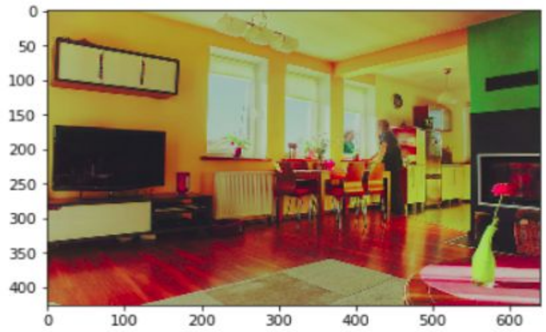
Loss : Cross-Entropy
Learning Rate : 0.001
Optimizer : Adam optimizer

Now , for a given unseen image final probability of each word of the dictionary with respect to that image is calculated in the following way.

$$P(word \mid image) = \sum_{i=1:K} (P(word \mid topic_i) P(topic_i \mid image))$$

Output Format

For an input image our model predicts ranks in accordance with decreasing probability. i.e. lower the probability higher the rank and rank 1 being the best.




```
(1, 2048)
Rank: 1 ROOM
Rank: 2 LIVE
Rank: 3 CHAIR
Rank: 4 COUCH
Rank: 5 STAND
Rank: 6 TABLE
Rank: 7 WOMAN
Rank: 8 LARGE
Rank: 9 ELEPHANT
Rank: 10 WALK
Rank: 11 FURNITURE
Rank: 12 WEAR
Rank: 13 ELEPHANTS
```

Results on unseen images.

Rank: 1 PIZZA
Rank: 11 SLICE

Rank: 2 SIGN
Rank: 6 STOP

Rank: 1 TENNIS
Rank: 2 BALL
Rank: 3 COURT
Rank: 4 PLAYER
Rank: 5 RACKET
Rank: 6 PLAY
Rank: 7 HOLD
Rank: 9 RACQUET
Rank: 10 SWING
Rank: 12 SERVE
Rank: 14 GAME
Rank: 15 MATCH
Rank: 18 MALE

Rank: 1 STAND
Rank: 2 FIELD
Rank: 3 GRASS
Rank: 9 GRASSY
Rank: 10 WALK
Rank: 11 GRAZE
Rank: 12 ZEBRA

Rank : 30 TOILET

Rank: 1 STREET
Rank: 2 SIGN
Rank: 28 CROSS

Rank: 1 TRAIN
Rank: 2 TRACK
Rank: 13 RAIL
Rank: 9 PLATFORM

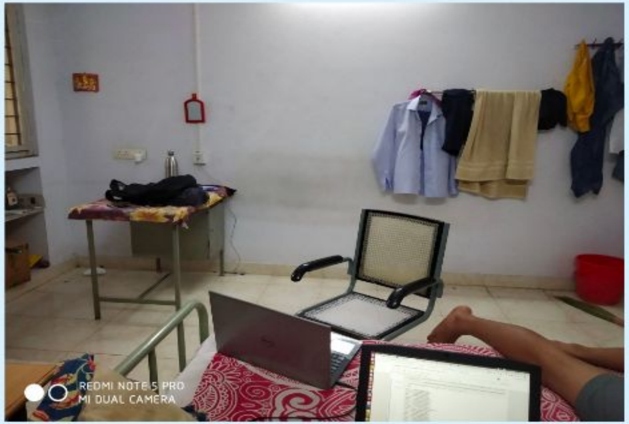
Rank: 161 SUBWAY
Rank: 3 FOOD
Rank: 7 SANDWICH
Rank: 18 VEGETABLES
Rank: 16 ENGINE

Rank: 39 STREET

Rank: 21 SMILE
Rank: 22 BABY
Rank: 8 CHILD
Rank: 2 WOMAN
Rank: 27 LADY
Rank: 72 FACE

Results of captured images

Rank: 2 HOLD
Rank: 4 TENNIS
Rank: 9 PERSON
Rank: 11 TABLE
Rank: 17 BALL
Rank: 25 PLAYER
Rank: 33 SERVE
Rank: 46 RACQUET



Rank: 1 LAPTOP
Rank: 2 DESK
Rank: 3 PEOPLE
Rank: 4 WINDOW
Rank: 7 TABLE
Rank: 14 SCREEN
Rank: 51 CHAIR
Rank: 68 GLASS
Rank: 171 SHIRT

Experiment we carried out !

We took 3 images having "leecooper" somewhere written on it. We manually added text of 4-5 lines for each image.

Testing for the image.

Rank : 91 LEECOOPER

Conclusion

Topic Modeling statistical framework can be used to leverage the correlation between visual and textual information in order to predict the words that are more likely to appear in the image as scene text instances. Moreover, we have shown that is possible to train a deep CNN model to reproduce those topic model based word rankings but using only an image as input.

Our Implementation shows that the quality of the automatically obtained custom lexicons is superior to a generic frequency based baseline, and thus can be used to improve scene text recognition methods.

Thank You