SMAI Project Report

DYNAMIC LEXICON GENERATION FOR NATURAL SCENE IMAGES

Interim Evaluation -II

Joyneel Misra 201401074
Sarthak Sharma 201431059
Shubham Rathore 201430101

INTRODUCTION

The underlying idea of our lexicon generation method is that the topic modeling statistical framework can be used to predict a ranking of the most probable words that may appear in a given image. For this we propose a three-fold method: First, we learn a LDA topic model on a text corpus associated with the image dataset. Second, we train a deep CNN model to generate LDA's topic-probabilities directly from the image pixels. Third, we use the generated topic-probabilities,either from the LDA model (using textual information) or from the CNN (using image pixels), along with the word-probabilities from the learned LDA model to re-rank the words of a given dictionary.

In the current interim evaluation,we solve the problem of generating a topic model of the document (text annotation + caption) associated with the images. Also we re-rank the words of the given dictionary given the topic-probabilities and word-probabilities from the underlying LDA model.

OVERVIEW

- Learning the LDA topic model using Textual Information.

Terms :

-*Document* : corresponds to the textual information associated to an image (e.g. image captions and scene text annotations).

-*Text corpus* : set of all textual information (documents) in the whole dataset.

The learned LDA model has two sets of parameters, the topic probabilities given documents

$$\mathbf{P(\text{topic} \mid \text{document})} \quad \text{and} \quad \mathbf{P(\text{word} \mid \text{topic})}$$

This way any new test document can be represented in terms of a probability distribution over topics of the learned LDA model.

- Using topic models for generating word ranks

Once the LDA topic model is learned as explained above, we can represent the document in terms of a probability distribution over topics. Also we already know the contribution of each word to each topic, P (word | topic) from the LDA model. We can calculate the probability of occurrence for each word in the dictionary **P (word | text)** as follows:

$$\mathbf{P(\text{word} \mid \text{text})} = \sum_{i=1}^{k} \mathbf{P(\text{word} \mid \text{topic i})} * \mathbf{P(\text{topic i} \mid \text{text})}$$

DATA

In our experiments we make use of two standard datasets, namely the MS-COCO and the COCO-Text datasets.

About the data set :

The MS-COCO is a large scale dataset providing task-specific annotations for object detection, segmentation, and image captioning. The dataset consists of 2.5 million labeled object instances among 80 categories in 328K images of complex everyday scenes. Images are annotated with multiple object instances and with 5 captions per image.

COCO-Text is a dataset for text detection and recognition in natural scene images that is based on the MS-COCO dataset. The images in this dataset were not taken with text in mind and thus it contains a broad variety of text instances. The dataset consists of 63,686 images, 173,589 text instances (words) and 3-fine grained text attributes. The dataset is divided in 43,686 training images and 20,000 validation images.

RESULTS

We show the result from the LDA for two sample images :

1.)

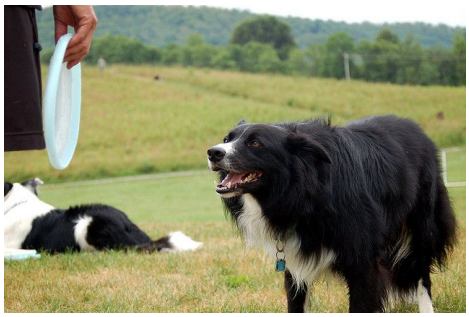
- Caption :

There is a robe and a towel folded on top of the bed", "Blanket at a hotel wrapped up on top of the bed", "A towel wrapped on top of a bed.", "A robe sitting on top of a white bed next to a wall.", "FOLDED ROBE TIED UP LIKE A PRESENT IN A HOTEL ROOM", "AND", "PACIFIC", "GR", "GRAND", "HOTEL", "HOTEL", "PAC"

- Output from LDA

room , baseball , kitchen , sit , live , table , chair , **bed** , couch , game , two , bat , white , plate , wii , small , next , large , floor , refriger , **top** , window , play , control , stove , televis , video , stand , food , train , player , remote , tv , man , wooden , counter , bedroom , furniture , fill , picture , area , oven , cabinet , **wall** , home , battery , red , front , black , hold , people , swing , book , open , bathroom , blue , look , mani , wood , woman , bowl , view , pot ,

2.)



- Caption:

The black and white dog stands near a person holding a Frisbee. ", "A dog is looking at a blue Frisbee.", "A dog watches a person who is holding a Frisbee.", "A dog looking at a man holding a Frisbee with another dog laying down.", "a couple of dogs that are in a grassy field"

- Output from LDA

dog , **man** , cat , water , **hold** , tennis , **stand** , beach , woman , sit , play , two , **lay** , boat , **person** , young , surfboard , ball , **black** , boy , white , wave , kite , player , field , ocean , girl , next , fli , people , **look** , **frisbee** , small , ride , surf , top , board , window , court , near , umbrella , bed , racket , group , little , men , larg , air , baseball , hit , hand , **blue** , surfer , brown , game , walk , child

REFERENCES

[1]

Dynamic Lexicon Generation for Natural Scene Images :Yash Patel[1,2] , Lluis Gomez[2] , Marçal Rusiñol[2] , and Dimosthenis Karatzas[2]