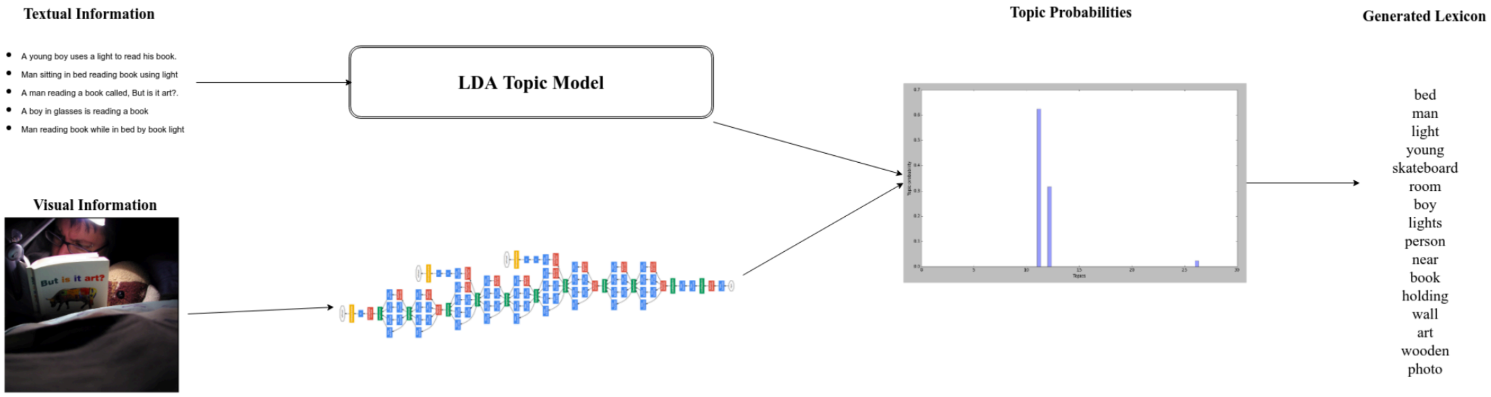
# Dynamic Lexicon Generation for Natural Scene Images

Joyneel Misra (201401074)
Sarthak Sharma (201431059)
Shubham Rathore (201430101)
IIIT Hyderabad



**Textual Information**

- A young boy uses a light to read his book.
- Man sitting in bed reading book using light
- A man reading a book called, But is it art?.
- A boy in glasses is reading a book
- Man reading book while in bed by book light

**LDA Topic Model**

**Visual Information**

**Topic Probabilities**

**Generated Lexicon**

bed
man
light
young
skateboard
room
boy
lights
person
near
book
holding
wall
art
wooden
photo

# Introduction

End-to-end scene text recognition pipelines are usually based in a multistage approach:

- First applying a text detection algorithm to the input image
- Then recognizing the text present in the cropped bounding boxes provided by the detector
- Scene text recognition from pre-segmented text is approached in two different conditions:
    - using a small provided lexicon per image
    - performing unconstrained text recognition, i.e. allowing the recognition of out-of-dictionary words.

The best performing end-to-end scene text understanding methodologies address the problem from a word spotting perspective and take a huge benefit from using customized lexicons. The size and quality of these custom lexicons has been shown to have a strong effect in the recognition performance, hence the motivation for the project.

The main intuition of the method is that visual information may provide in some cases a valuable cue for text recognition algorithms: there are some words for which occurrence in a natural scene image correlates directly with objects appearing in the image or with the scene category itself.

In this project, we implement a method that generates contextualized lexicons based only on visual information. For this we make the following contributions:

- We learn a topic model using Latent Dirichlet Allocation (LDA) using as a corpus textual information associated with scene images combined with scene text.
    - This LDA model is suitable for generating contextualized lexicons of scene text given image descriptions by reranking the words in the text.

- We train a deep CNN model, based on the LDA topic model, that is capable to produce on its output a topic probability distribution per document given by the LDA analysis directly from the corresponding image input.
- 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.

We also show that the quality of such automatically obtained custom lexicons is superior to generic frequency-based lexicons in predicting the words that are more likely to appear as scene text instances in a given image.

# Method

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.

## Learning the LDA topic model using Textual Information

Our method assumes that the textual information associated with the images in our dataset is generated by a mixture of latent topics. LDA is a generative statistical model of a corpus (a set of text documents) where each document can be viewed as a mixture of various topics, and each topic is characterized by a probability distribution over words. LDA can be represented as a three level hierarchical Bayesian model.

- 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 and the word probabilities given topics.

$$P(\,topic \mid document\,) \quad and \quad P(\,word \mid topic\,)$$

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

## Training a CNN to predict probability distributions over LDA's topics

Once we have the LDA topic model, we want to train a deep CNN model to predict the same probability distributions over topics as the LDA model does for textual information, but using only the raw pixels of new unseen images. For this we can generate a set of training (and validation) samples as follows: given an image from the training set we represent its corresponding textual information (captions) as probability values over the LDA's topics. These probability values are used as labels for the given image.

This way we obtain a set of M training (and validation) examples of the form{(x1,y1),..., (xM,yM)}such that xi is an image and yi is the probability distribution over topics obtained by projecting its associated textual information into the LDA topic space.

Using this training set we train a deep CNN to predict the probability distribution yi for unseen images. In fact, we use a transfer learning approach here in order to shortcut the training process by fine-tuning the well known Inception deep CNN 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:

$$P(word \mid text) = \Sigma\, P(word \mid topic\ i) * P(topic\ i \mid text)$$

Similarly, once the deep CNN is trained using the LDA model, we can obtain the probability distribution over topics for an unseen image P (topic | image) as the output of the CNN when feeding the image pixels on its input. Again, we can calculate the probability of occurrence of each word in the dictionary P (word | image) as follows :

$$P(word \mid image) = \Sigma\, P(word \mid topic\ i) * P(topic\ i \mid image)$$

Using the obtained probability distributions over words, we are able to rank a given dictionary in order to prioritize the words that have more chances to appear in a given image.

# Experiments and Results

In this section we present the experimental evaluation of the proposed method on its ability to generate lexicons that can be used to improve the performance of systems for reading text in natural scene images.

## Datasets

- Corpus - We learned the LDA topic model using only the 43686 images in the train set of COCO-Text and their corresponding captions (from MS-COCO)
- Dictionary - We do experiments with two different dictionaries:
  - The list of 15183 (stopped and stemmed) unique annotated text instances (words) in the COCO-Text dataset.
  - A generic dictionary of approximately 88172 words used, but we removed stop words and stemmed words thus giving rise to a dictionary of 87855 words.

## Implementation Details

In our experiments involving topic modeling we have used the gensim  Python library for learning and inferring the LDA model. We have learned multiple LDA models with a varying number of topics and have compared word ranking results.

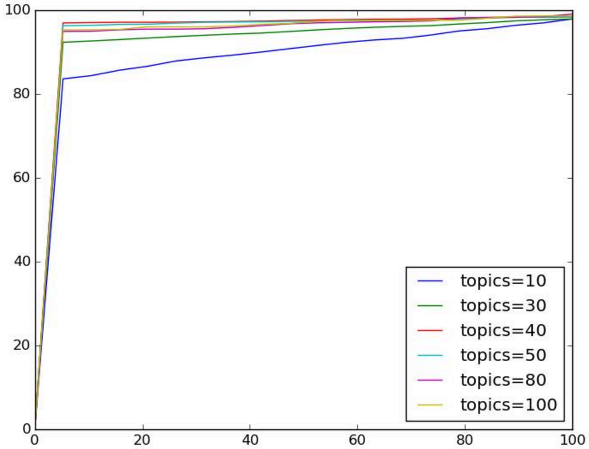
On the other hand, we have used the TensorFlow  framework for fine- tuning of the Inception v3 model . We have trained the final layer of the net from scratch, accommodating it to the size of our topic modeling task, and leaving the rest of the net untouched. We used the cross entropy loss function and Gradient Descent optimizer with a fixed learning rate of 0.01 and a batch size of 100 for 100k iterations.

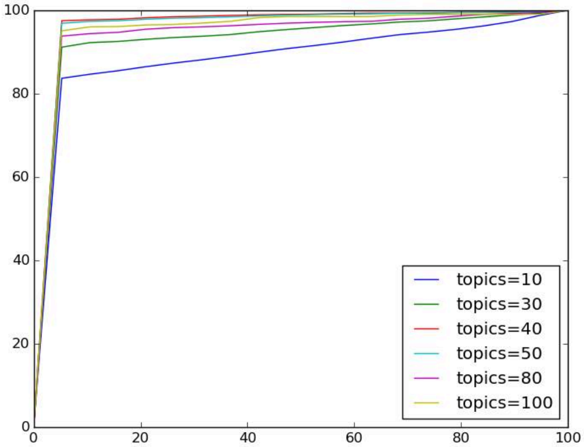
## Word-rankings using LDA

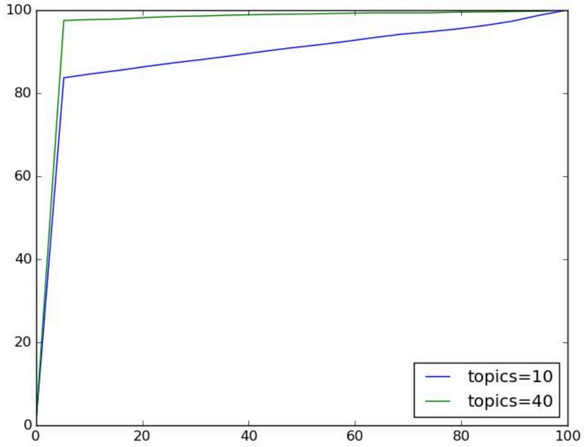
Given a fixed dictionary we are interested in word rankings (lexicons generated) that are able to prioritize the words that are more likely to appear in a given image as scene text instances. Thus, we use the following procedure to evaluate and compare different word rankings:

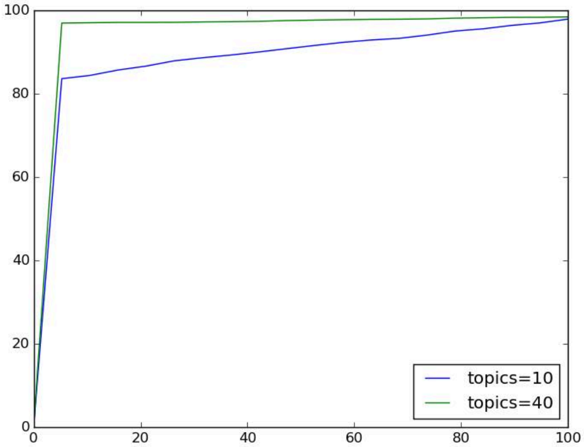
- For every word ranking we count the percentage of COCO-Text ground truth (validation) instances that are found among the top-N words of the re-ranked dictionary.

This way we can plot curves illustrating the number of COCO-Text instances found in different word lists (lexicons) that correspond to certain top percentages of the ranked dictionary. The larger the area under those curves the better is a given ranking.









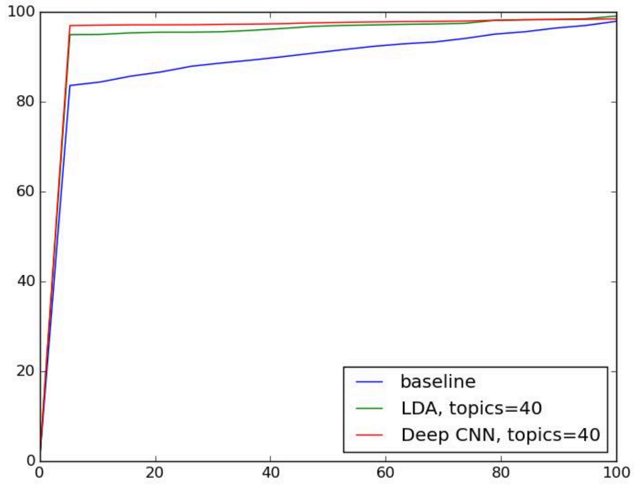
X axis : Percentage of words in (re-ranked) dictionary, Y axis : Percentage of COCO-Text instances found. Figures above ( clockwise from top-left)

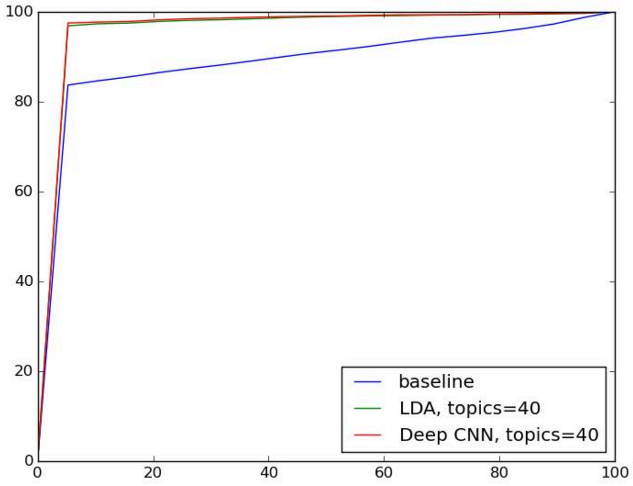
- Fig-1 : Corpus with Dictionary(2) by varying the number of topics between 10-100
- Fig-2 : Corpus with Dictionary(1) by varying the number of topics between 10-100
- Fig-3 : Corpus with Dictionary(2) (topics=40) vs Baseline
- Fig-4 : Corpus with Dictionary(1) (topics=40) vs Baseline

The number of topics in the LDA topic model is an important parameter of the method. The best performance for our automatically generated rankings are num_topics = 40. In such a case the performance of the LDA based rankings is superior to the baseline in all the experiments.

- This demonstrates that the LDA topic modeling analysis we implement predicts the occurrence of words as scene text instances better than a frequency-based dictionary.

## Word-rankings using CNN

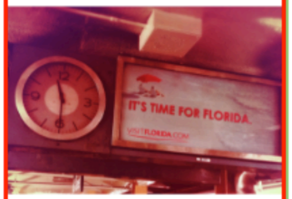




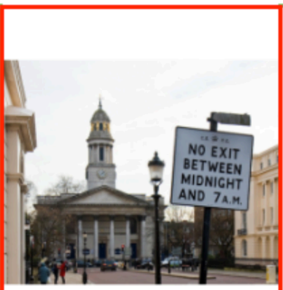
The figures above show the performance comparison of the word rankings obtained by the LDA model using 40 topics as in the previous section, and the word rankings obtained with the CNN. We can see that the CNN is able to produce word rankings with almost the same performance as projecting the images' captions in the LDA space, but using only the image raw pixels as input. Using the CNN for predicting the probability
distribution over 40 topics for a given image takes takes less than 0.1s.
As shown in the figure above, performance of the word rankings obtained directly from the topic model and the CNN are almost identical, we can conclude that these values are only an estimator of the CNN real performance.

## Conclusion

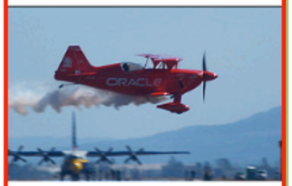
In this paper we have presented a method that generates automatic contextualized per image lexicons based on visual information using deep CNN and LDA topic model. This way we make use of the rich visual information contained in scene images that could provide help to improve text detection and recognition results. We have also shown that is possible to train a deep CNN model to reproduce the LDA topic model based word rankings but using only an image as input.



**Word annotations :**
**Word Rank**
florida : 7919
time : 167
**Top Ranked words**
clock : 127
red : 5
front : 13
restaurant : 213

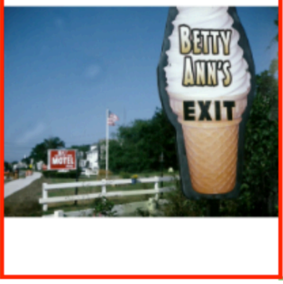


**Word annotations :**
**Word Rank**
midnight : 13900
exit : 2041
**Top Ranked words**
sign : 2
building : 3
street : 4
road : 29



**Word annotations :**
**Word Rank**
oracle : 8488

**Top Ranked words**
flying : 31
top : 33
airplanes : 344
motor : 314



**Word annotations :**
**Word Rank**
betty : 71215
exit : 1534
**Top Ranked words**
street : 5
flag : 354
sky : 104
fence : 112

# References

- Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan) (2003) 993–1022

- Patel, Y., and Gomez, L., and Rusiñol, M., and Karatzas, D.: Dynamic Lexicon Generation for Natural Scene Images

- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.00567 (2015)