Dynamic Lexicon Generation for Natural Scene Images

Joyneel Misra
Sarthak Sharma
Shubham Rathore

Mentor: Yash Patel

Introduction

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

- Text detection algorithm to the input image.
- Recognition of the text present
    - Text present in the cropped bounding boxes provided by the detector
    - Scene text recognition from pre-segmented text approached in two ways:
        i. using a small provided lexicon per image
        ii. performing unconstrained text recognition (allowing the recognition of out-of-dictionary words).

Introduction

Current methods take a huge benefit from using customized lexicons.

Motivation : The size and quality of these custom lexicons has been shown to have a strong effect in the recognition performance of text in natural scene images.

Intuition : Words occurring in a natural scene image correlate directly with objects appearing in the image or with the scene category itself. This implies visual information may provide in some cases a valuable cue for text recognition algorithms.

Method

We implement a method that generates contextualized lexicons based only on visual information. Method includes :

- Learning a topic model using Latent Dirichlet Allocation (LDA) using as a corpus textual information associated with scene images combined with scene text.
- Train a deep CNN model, based on the LDA topic model.
- Generate contextualized lexicons for new (unseen) images directly from their raw pixels, without the need of any associated textual content.

Learning the LDA topic model

The method assumes that the textual information (image descriptions + captions)associated with the images in the 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.

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

P(topic | document) and P(word | topic)

Training CNN based on LDA topic model

We 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 images in the dataset.

Input for training **CNN** -------> Image + Topic probability per doc[from **LDA**]

Topic probability per document is used as a target output for the CNN

The LDA model gives us the topic probability distribution of the document (textual information of an image).Similarly, the trained CNN predicts the topic probability distribution using only the images as it's input.

Using topic models to generate word ranks

Topic model is learned as discussed in previous slides,

- We can represent the image in terms of a probability distribution over topics (from the learnt CNN). 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 | image) as follows:

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

- We are able to rank a given dictionary in order to prioritize the words that have more chances to appear in a given image.

Frameworks and Implementation details

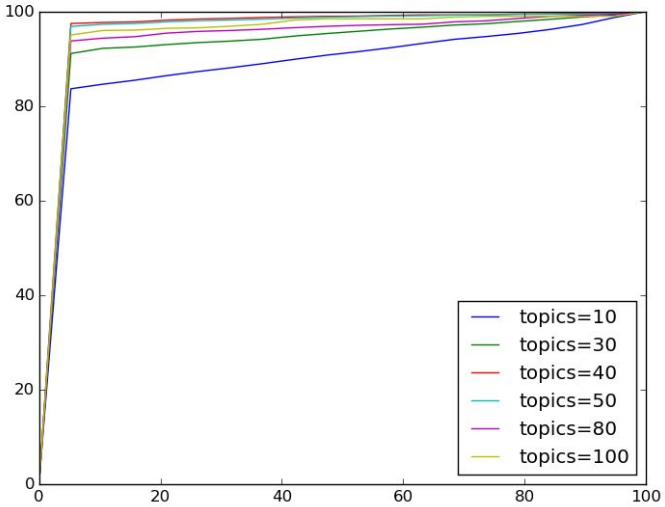
LDA : Gensim

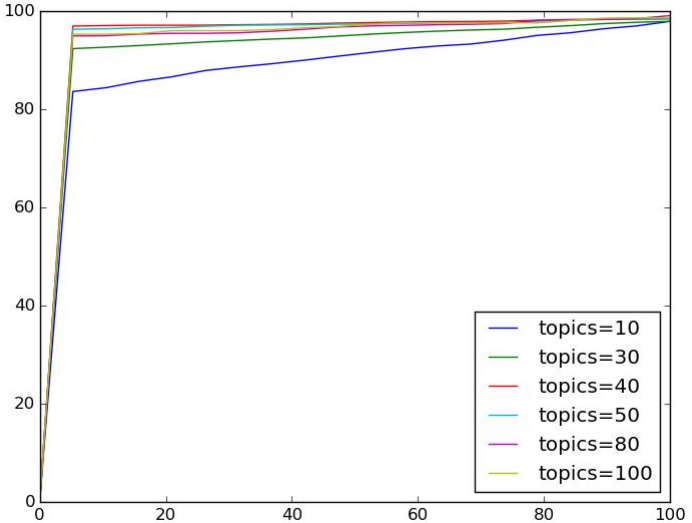
CNN : Tensor Flow (Fine tuning Inception_v3 model)

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)

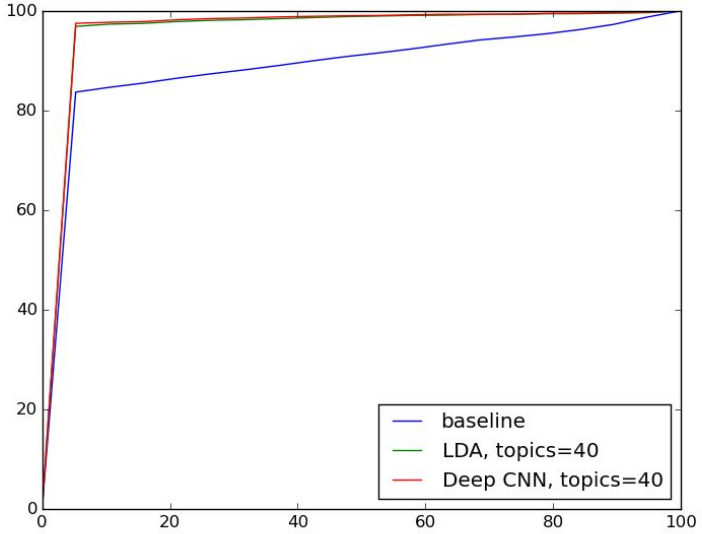
Dictionary : We do experiments with two different dictionaries. (a) The list of 15183 (stopped and stemmed) unique annotated text instances (words) in the COCO-Text dataset. (b) A generic dictionary (stopped and stemmed) of approximately 87855 words used.

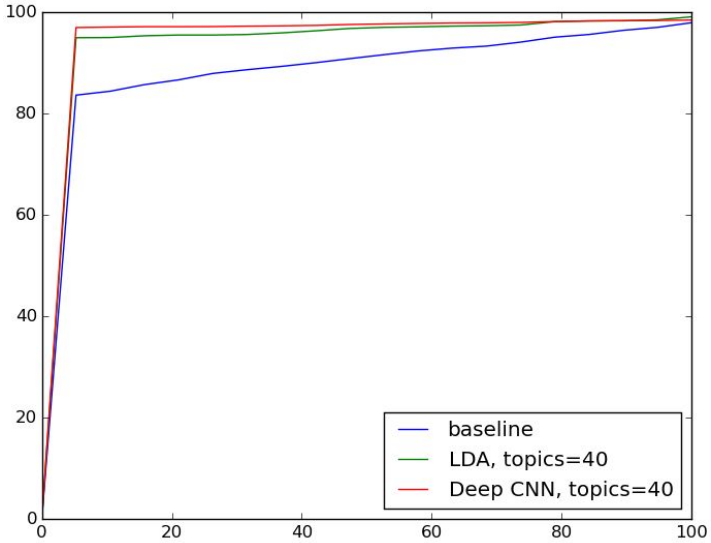
Results (word rankings using LDA)





Results (word rankings using CNN)





Conclusion

- A method that generates automatic contextualized per image lexicons based on visual information using deep CNN and LDA topic model.
- The method makes use of the rich visual information contained in scene images (according to the intuition) 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.