ITMO UNIVERSITY

Multimodal topic model for texts and images utilizing their embeddings

Nikolay Smelik, smelik@rain.ifmo.ru

Andrey Filchenkov, afilchenkov@corp.ifmo.ru

Computer Technologies Lab

IDP-16.
Barcelona, Spain, 10–14 Oct 2016

Topic model

- Topic model is defined as a model of document collection that put into correspondence each document in the collection and a topic.
- A **document** is not only a textual document, but any structured object represented by a set of elements, such as, for instance, user described with preferences.

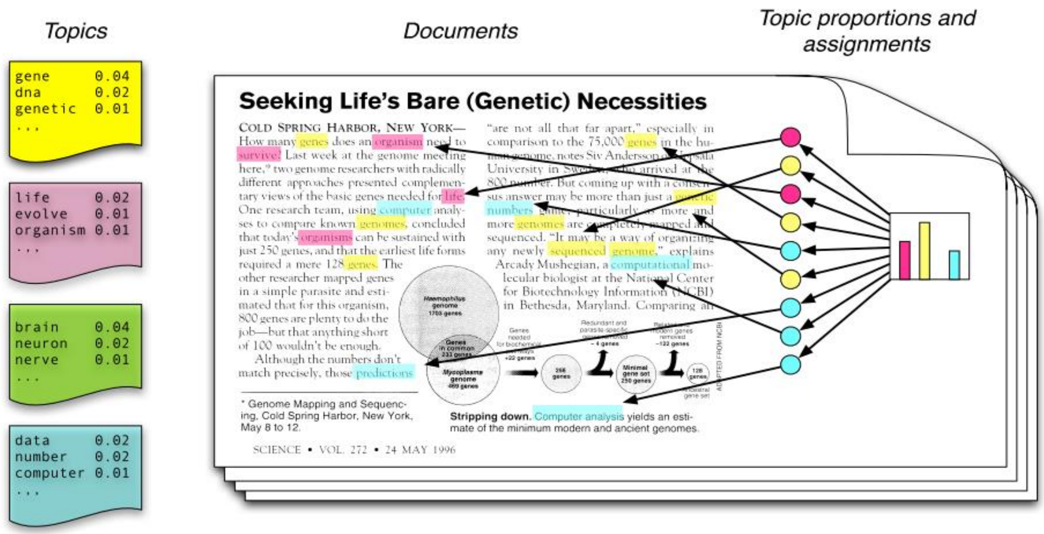

Illustration is taken from Blei, D. (2012) Probabilistic Topic Models

Multimodal topic models for text and images

Multimodal topic models for text and images are useful for

- Text annotating and image illustrating
- Text search given image and image search given description
- Topic clustering
- Image synthesis from textual description
- ...

Research goal

The goal of this research is to increase quality of image annotating and search given textual description.

We achieve this goal by creating a multimodal topic model that utilize

- word embedding;
- convolutional neural network image representation.

Related works

- CorLDA[1]
  - Image feature are extracted from segments
  - Correspondence LDA as a topic model
- MixLDA[2]
  - Image feature are extraction with SIFT and clustered
  - Multimodal LDA as a topic model
- sLDA[3]
  - Image feature are extraction with SIFT and clustered
  - Supervised LDA as a topic model

[1]BleiD. M., Jordan M.I. (2003) Modeling annotated data // SIGIR. 127-134.
[2]Feng Y., Lapata M. (2010) Topic models for image annotation and text illustration // Human Language Technologies. 831-839.
[3]Wang C., Blei D., Li F.F. (2009) Simultaneous image classification and annotation // CVPR. 1903-1910.

Steps to build the model

- Image preprocessing
- Last convolutional layer vector extraction (*image vector*)
- Text preprocessing
- Word embedding vector extraction (*word vector*)
- Topic model learning given set of image vectors described by word vectors

Topic model learning step

- Image vector $\mathbf{i}$ is considered as a pseudo document, in which "words" are word vectors $\mathbf{w}$ of words from the image description
- TF-IDF is used to describe probability of "words" in a psedo-document
- A matrix $F = (p_{\mathbf{wi}})_{|W| \times |I|}$ is build, where $p_{\mathbf{wi}} = tfidf(\mathbf{w}, \mathbf{i}, I)$
- *F* is represented as a multiplication of matrices Φ and Θ:

$$F \approx \Phi\Theta.$$

Topic model learning step: F decomposition (1/2)

Representing F as $F \approx \Phi\Theta$ is matrix decomposition, where

- Φ represents conditional distributions on “words” given topic
- Θ represents conditional distributions on topics given images

This problem is solved by maximizing log-likelihood given constraints on normalization and non-negativity of rows:

$$L(\Phi, \Theta) = \sum_{\mathbf{i} \in I} \sum_{\mathbf{w} \in \mathbf{i}} p_{\mathbf{wi}} \ln \sum_{t \in T} \varphi_{\mathbf{w}t} \theta_{t\mathbf{i}} \to \max_{\Phi, \Theta};$$

$$\sum_{\mathbf{w} \in W} \varphi_{\mathbf{w}t} = 1; \ \varphi_{\mathbf{w}t} \geq 0; \ \sum_{t \in T} \theta_{t\mathbf{i}} = 1; \ \theta_{ti} \geq 0.$$

Topic model learning step: F decomposition (2/2)

The problem met by the described approach is that there are many (maybe, infinite number of) solutions to $F \approx \Phi\Theta$.

This problem can be handled by adding model regularization for Θ and Φ: $R_i(\Phi, \Theta)$.

The problem is thus reduced to the problem of maximizing a linear combination of L and all R_i under the same restrictions:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta),$$

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Scheme for image annotating

- Input is an image
- Image vector $\mathbf{i}_{input}$ is evaluated
- A closest image vector $\mathbf{i}_{NN}$ from all the known image vectors is found
- Most probable topics are evaluated for $\mathbf{i}_{NN}$
- Words, relevant to the found topics are extracted
- Output are these words

Scheme for image search given annotations

- Input is an annotation
- Word vector $\mathbf{w}_{input}$ is evaluated
- A closest word vector $\mathbf{w}_{NN}$ from all the known word vectors is found
- Most probable topics are evaluated for $\mathbf{w}_{NN}$
- Image vectors, relevant to the found topics are extracted
- Output are these images

Dataset

- Microsoft Common Object in Context
- 21000 images
- At least five annotation for each image
- Vocabulary size is 6000

Dataset

- Microsoft Common Object in Context
- 21000 images
- At least five annotation for each image
- Vocabulary size is 6000

Example

- A cat sits on the edge of a bathroom sink.
- A grey tabby cat sitting on the sink in a bathroom.
- A cat sitting up on a counter in a room.
- A cat is sitting perched on the corner of a bathroom sink.
- A grey and black cat sitting next to sink in a bathroom.

Dataset

- Microsoft Common Object in Context
- 21000 images
- At least five annotation for each image
- Vocabulary size is 6000

Image preprocessing step

- Convolutional neural network without fully-connected layers
- We used a learnt network VGG-16
- All the images are compressed to size 224×224
- As a result, we obtain an image vector

Text preprocessing step

- All symbols that are not letters/digits are filtered out
- Words are extracted
- Stop-words are eliminated
- Normal form of words is used
- A learnt Word2Vec from Gensim library is used
- For each word, word vector is obtained.

Model quality evaluation

- Perplexity: $P = \exp\left(-\frac{1}{n}\sum_{\mathbf{i}\in I}\sum_{\mathbf{w}\in i} n_{\mathbf{iw}} \ln p(\mathbf{w}|\mathbf{i})\right)$;
- Matrices sparsity S_Φ and S_Θ;
- Purity: $K_\mathrm{p} = \sum_{\mathbf{w}\in W_t} p(\mathbf{w}|t)$;
- Contrast: $K_\mathrm{c} = \frac{1}{|W_t|}\sum_{\mathbf{w}\in W_t} p(t|\mathbf{w})$.

Model	P	S_Φ	S_Θ	K_p	K_c
ARTM	**70.312**	**96.5**	**88.6**	**0.889**	**0.831**
PLSA	84.596	82.1	84.6	0.461	0.656

Image annotating results

- Data: MS COCO dataset
- Data split: 80/20
- Method of evaluation: comparison of top 10 words predicted by a model with top 10 words from description ranked with tf

Model	Recall	Precision	F_1-measure
CorrLDA	34.83	37.85	36.27
MixLDA	35.20	37.98	36.54
sLDA	35.63	38.46	36.99
PLSA	35.94	38.02	36.92
ARTM	**40.43**	**43.37**	**41.85**

Example of image annotating

Model		
CorrLDA	**motorcycle**, **road**, motor, **street**, bike, **car**, subway, garage, building	**train**, **platform**, tree, sky, car, building, subway, lake, house, person
MixLDA	**motorcycle**, person, **jacket**, **street**, helmet, **crossroad**, parked, tree, water	**train**, **passenger**, **hill**, road, **track**, black, sky, window, tree, forest
sLDA	**motorcycle**, **road**, **jacket**, **street**, parked, garage, sidewalk, bike, helmet	**train**, **platform**, **passenger**, building, car, street, sky, **hill**, **track**, person
PLSA	**motorcycle**, **road**, motor, **street**, **biker**, **crossroad**, parked, garage, tire	**train**, **station**, **track**, **platform**, **passenger**, engine, menu, subway, vase
ARTM	**motorcycle**, **road**, **car**, **street**, **biker**, **crossroad**, parked, garage, sidewalk	**train**, **station**, **track**, **platform**, **passenger**, engine, **hill**, road, tree
Real description	motorcycle, bikers, street, car, traffic, jacket, crossroad, road	train, track, station, hill, platform, standing, passenger, hat, steam

Image search results

- Data: MS COCO dataset
- Image pool size is 5
- Method of evaluation: percentage of found pairs (image – description)

Model	Accuracy
MixLDA	43.5
PLSA	55.8
ARTM	**60.4**

Image search example

Input description: *Men wearing baseball equipment on a baseball field*

Image search example

Input description: *Men wearing baseball equipment on a baseball field*





Conclusion

- We suggested a multimodal topic model for texts and images utilizing CNNs and Word2Vec
- We showed that the model outperforms state-of-the-art approaches in image annotating and image search
- Our future work is no use word vectors and image vectors as frequency vectors and build a topic model for “contexts”

ITMO UNIVERSITY

Thank you!

Multimodal topic model for texts and images utilizing their embeddings

Nikolay Smelik and Andrey Filchenkov,

smelik@rain.ifmo.ru, afilchenkov@corp.ifmo.ru