aletras_2017_labeling_topics_with_images_using_a_neural_network

Year

2017

Author(s)

Aletras, Nikolaos and Mittal, Arpit

Title

Labeling Topics with Images Using a Neural Network

Venue

ECIR

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel

Underlying technique

Deep Neural Network

Topic labeling parameters

- Terms used to represent the topic (t): 10
- Embedding size (d): 300
- Hidden layer regularisation: Dropout (Srivastava et al., 2014)

- Output size of hidden layers: 256, 128, 64 and 32
- Loss function: mean absolute error (MAE)
- Optimizer: mini-batch gradient descent method with RMSProp adaptive learning rate algorithm
- Dropout value: 0.2 (randomly sets 20% of the input units to 0 at each update during the training time)
- Nr folds (k-fold cross-validation): 5
- Epochs: 30
- Batch size for training data: 16

Label generation

A Deep Neural Network (DNN) is utilised to estimate the suitability of an image for labeling a given topic.
For a topic T and an image I, we want to compute a real value s ∈ R that denotes how good the image I is for representing the topic T.
T consists of ten terms (t) with the highest probability for the topic.
We denote the visual information of the image as V .
The image is also associated with text in its caption, C.
For the topic T = {t1, t2, …, t10} and the image caption C = {c1, c2, …, cn}, each term is transformed into a vector

$$\mathbf{x} \in \mathbb{R}^d$$

where d is the dimensionality of the distributed semantic space.
We use pre-computed dependency-based word embeddings (Levy and Goldberg, 2014) whose d is 300.
The resulting representations of T and C are the mean vectors of their constituent words, xt and xc respectively.

The visual information from the image V is converted into a dense vectorised representation, xv.
That is the output of the publicly available 16-layer VGG-net Simonyan and Zisserman, 2014 trained over the ImageNet dataset (Deng et al., 2009).
VGG-net provides a 1000 dimensional vector which is the soft-max classification output of ImageNet classes.
The input to the network is the concatenation of topic, caption and visual vectors. i.e.,

$$X = [x_t || x_c || x_v]$$

This results in a 1600-dimensional input vector.

Then, X is passed through a series of four hidden layers, H1,…,H4.
In this way the network learns a combined representation of topics and images and the non-linear relationships that they share.

$$h_i = g(W_i^T h_{i-1})$$

where g is the rectified linear unit (ReLU) and h0 = X.
The output of each hidden layer is regularised using dropout.
The output size of H1,H2,H3 and H4 are set to 256, 128, 64 and 32 nodes respectively.
The output layer of the network maps the input to a real value s ∈ R that denotes how good the image I is for the topic T.
The network is trained by minimising the mean absolute error:

$$error = \frac{1}{n}\sum_{i=1}^{n} \left| W_o^T h_4 - s_g \right|$$

where sg is the ground-truth relevance value.
The network is optimised using a standard mini-batch gradient descent method with RMSProp adaptive learning rate algorithm

In each fold, data from 240 topics are used for training which results into 9,600 examples (20 original, 20 negative candidates per topic).
The rest completely unseen 60 topics are used for testing which results into 1,200 test examples
(note that we do not add negative examples in the test data).

Motivation

Offering a language independent representation of the topic which can also be complementary to textual labels.
The visual representation of a topic has been shown to be as effective as the textual labels on retrieving information using a topic browser while it can be understood quickly by the users (Aletras et al., 2014, Alteras et al., 2017)

Topic modeling

LDA

Topic modeling parameters

Nr of topics (K): 200, 400

α: 1/num_topics

β: 1/num_topics

Nr. of topics

600 (total)

Label

Set of images associated by the DNN to the top 10 words of a topic

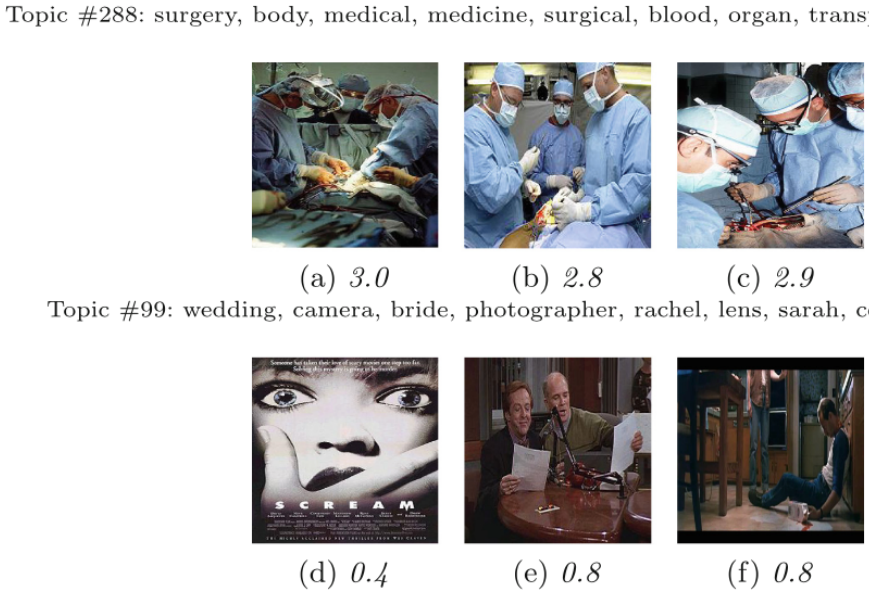

Fig. 1. A good and a bad example of topics and the top-3 images (left-to-right) selected by the DNN (Topic+Caption+VGG) model from the candidate set. Subcaptions denote average human ratings.

Label selection

\

Label quality evaluation

- The Top-1 average rating: Average human rating assigned to the top-ranked label proposed by the topic labeling method. This metric provides an indication of the overall quality of the label selected and takes values from 0 (irrelevant) to 3 (relevant).
- The normalized discounted cumulative gain (nDCG): Compares the label ranking proposed by the labeling method to the gold-standard ranking provided by the human annotators.

Comparison models

The approach is compared with:

- Local PPR: state-of-the-art method that uses Personalized PageRank to re-rank image candidates (Aletras and Stevenson, 2013)
- Global PPR: an adapted version that computes the PageRank scores of all the available images in the test set.
- WSABIE: A relevant approach originally proposed for image annotation that learns a joint model of text and image features (Weston et al., 2010)
- LR (Topic+Caption+VGG), SVM (Topic+Caption+VGG): linear regression and SVM models that use the concatenation of the topic, the caption and the image vectors as input.
- DNN (Topic+Caption), DNN (Topic+VGG): Two versions of the proposed DNN using only either the caption or the visual information of the image.

Table 1. Results obtained for the various topic labeling methods. †, ‡ and ∗ denote statistically significant difference to Local PPR, Global PRR and WSABIE respectively (paired t-test, $p < 0.01$).

Model	Top-1 aver. rating	nDCG-1	nDCG-3	nDCG-5
Global PPR [3]	1.89	0.71	0.74	0.75
Local PPR [3]	2.00	0.74	0.75	0.76
WSABIE [20]	1.87	0.65	0.68	0.70
LR (Topic+Caption+VGG)	1.91	0.71	0.74	0.75
SVM (Topic+Caption+VGG)	1.94	0.72	0.75	0.76
DNN (Topic+Caption)	1.94	0.73	0.75	0.76
DNN (Topic+VGG)	$2.04^{\ddagger *}$	0.76	0.79	0.80
DNN (Topic+Caption+VGG)	$\mathbf{2.12}^{\dagger\ddagger *}$	**0.79**	**0.80**	**0.81**
Human Perf. [3]	2.24	-	-	-

Assessors

\

Domain

Domain (paper): Topic labeling
Domain (dataset): Miscellaneous (Wikipedia), news

Problem statement

Presenting a more generic method that can estimate the degree of association between any arbitrary pair of an unseen topic and image using a deep neural network.

Corpus

Origin: Wikipedia and news articles
Content: 300 topics with associated candidate image labels and their human ratings
Details: Dataset generated in Aletras and Stevenson, 2013

Document

A topic, represented by 10 terms and associated with 20 candidate image labels and their human ratings between 0 (lowest) and 3 (highest) denoting the appropriateness of these images for the topic (resulting in a total of 6k images and associated captions).

Pre-processing

The 20 candidate image labels per topic are collected by using Google. Hence most of them are expected to be relevant to the topic.
To add sufficient negative examples, another 20 images are sampled for each topic from random topics in the training set and are assigned a relevance score of 0. These extra images are added into the training data.

```
@inproceedings{aletras_2017_labeling_topics_with_images_using_a_neural_network,
  abstract = {Topics generated by topic models are usually represented by lists
```

```
of t terms or alternatively using short phrases or images. The current state-of-the-art work on labeling topics using images selects images by re-ranking a small set of candidates for a given topic. In this paper, we present a more generic method that can estimate the degree of association between any arbitrary pair of an unseen topic and image using a deep neural network. Our method achieves better runtime performance O(n) compared to {\$}{\$}O(n^2){\$}{\$}for the current state-of-the-art method, and is also significantly more accurate.},
  address = {Cham},
  author = {Aletras, Nikolaos and Mittal, Arpit},
  booktitle = {Advances in Information Retrieval},
  date-added = {2023-03-03 17:06:21 +0100},
  date-modified = {2023-03-03 17:06:21 +0100},
  editor = {Jose, Joemon M and Hauff, Claudia and Alt{\i}ngovde, Ismail Sengor and Song, Dawei and Albakour, Dyaa and Watt, Stuart and Tait, John},
  isbn = {978-3-319-56608-5},
  pages = {500--505},
  publisher = {Springer International Publishing},
  title = {Labeling Topics with Images Using a Neural Network},
  year = {2017}}
```

#Thesis/Papers/Initial