# Labeling Topics with Images Using a Neural Network

Nikolaos Aletras(✉) and Arpit Mittal

Amazon.com, Cambridge, UK
{aletras,mitarpit}@amazon.com

**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.

**Keywords:** Topic models · Deep neural networks · Topic representation

## 1 Introduction

Topic models [5] are a popular method for organizing and interpreting large document collections by grouping documents into various thematic subjects (e.g. sports, politics or lifestyle) called topics. Topics are multinomial distributions over a predefined vocabulary whereas documents are represented as probability distributions over topics. Topic models have proven to be an elegant way to build exploratory interfaces (i.e. topic browsers) for visualizing document collections by presenting to the users lists of topics [6,14,15] where they select documents of a particular topic of interest.

A topic is traditionally represented by a list of $t$ terms with the highest probability. In recent works, short phrases [4,11], images [3] or summaries [19] have been used as alternatives. Particularly, images offer 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 [1,2]. The task of labeling topics consists of two main components: (1) a candidate generation component where candidate labels are obtained for a given topic (usually using information retrieval techniques and knowledge bases [3,11]), and (2) a ranking (or label selection) component that scores the candidates according to their relevance to the topic. In the case of labeling topics with images the candidate labels consist of images.


J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 500–505, 2017.
DOI: 10.1007/978-3-319-56608-5_40

The method presented by [3] generates a graph where the candidate images are its nodes. The edges are weighted with a similarity score between the images that connect. Then, an image is selected by re-ranking the candidates using PageRank. The method is iterative and has a runtime complexity of $O(n^2)$ which makes it infeasible to run over large number of images. Hence, for efficiency the candidate images are selected a priori using an information retrieval engine. Thus the scope of this method gets limited to solving a *local problem* of re-ordering a small set of candidate images for a given topic. Furthermore, its accuracy is limited by the recall of the information retrieval engine. Finally, if new candidates appear, they should be added to the graph, the process of computing pairwise similarities and re-ranking of nodes is repeated.

In this work, we present a more generic method that directly estimates the appropriateness of any arbitrary pair of topic and image. We refer to this method as a *global method* to differentiate it from the localized approach described above. We utilize a Deep Neural Network (DNN) to estimate the suitability of an image for labeling a given topic. DNNs have proven to be effective in various IR and NLP tasks [7,16]. They combine multiple layers that perform non-linear transformations to the data allowing the automatic learning of high-level abstractions. At runtime our method computes dot products between various features and the model weights to obtain the relevance score, that gives it an order complexity of $O(n)$. Hence, it is suitable for using it over large image sources such as Flickr[1], Getty[2] or ImageNet [9]. The proposed model obtains state-of-the-art results for labeling completely unseen topics with images compared to previous methods and strong baselines.

## 2 Model

For a topic $T$ and an image $I$, we want to compute a real value $s \in \mathbb{R}$ 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 $T = \{t_1, t_2, ..., t_{10}\}$ and the image caption $C = \{c_1, c_2, ..., c_n\}$, 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 [12] whose $d$ is 300. The resulting representations of $T$ and $C$ are the mean vectors of their constituent words, $\mathbf{x_t}$ and $\mathbf{x_c}$ respectively.

The visual information from the image $V$ is converted into a dense vectorized representation, $\mathbf{x_v}$. That is the output of the publicly available 16-layer VGG-net [13] trained over the ImageNet dataset [9]. 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.,

$$X = [x_t||x_c||x_v] \tag{1}$$

This results in a 1600-dimensional input vector.

[1] http://www.flickr.com.
[2] http://www.gettyimages.co.uk.

Then, $X$ is passed through a series of four hidden layers, $H_1, ..., H_4$. In this way the network learns a combined representation of topics and images and the non-linear relationships that they share.

$$h_i = g(W_i^T h_{i-1}) \tag{2}$$

where $g$ is the rectified linear unit (ReLU) and $h_0 = X$. The output of each hidden layer is regularized using dropout [17]. The output size of $H_1, H_2, H_3$ and $H_4$ are set to 256, 128, 64 and 32 nodes respectively.

The output layer of the network maps the input to a real value $s \in \mathbb{R}$ that denotes how good the image $I$ is for the topic $T$. The network is trained by minimizing the mean absolute error:

$$error = \frac{1}{n} \sum_{i=1}^{n} \left| W_o^T h_4 - s_g \right| \tag{3}$$

where $s_g$ is the ground-truth relevance value. The network is optimized using a standard mini-batch gradient descent method with RMSProp adaptive learning rate algorithm [18].

## 3 Experimental Setup

We evaluate our model on the publicly available data set provided by [3]. It consists of 300 topics generated using Wikipedia articles and news articles taken from the New York Times. Each topic is represented by ten terms with the highest probability. They are also 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. That results into a total of 6 K images and their associated textual metadata which are considered as captions. The task is to choose the image with the highest rating from the set of the 20 candidates for a given topic.

The 20 candidate image labels per topic are collected by [3] using an information retrieval engine (Google). Hence most of them are expected to be relevant to the topic. This jeopardizes the training of our supervised model due to the lack of sufficient negative examples. To address this issue we generate extra negative examples. For each topic we sample another 20 images from random topics in the training set and assign them a relevance score of 0. These extra images are added into the training data.

Our evaluation follows prior work [3,11] using two metrics. The **Top-1 average rating** is the 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 [8,10].

We set the dropout value to 0.2 which randomly sets 20% of the input units to 0 at each update during the training time. We train the model in a 5-fold cross-validation for 30 epochs and set the batch size for training data to 16. 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).

## 4 Results and Discussion

We compare our approach to the state-of-the-art method that uses Personalized PageRank [3] to re-rank image candidates (**Local PPR**) and an adapted version that computes the PageRank scores of all the available images in the test set (**Global PPR**). We also test other baselines methods: (1) a relevant approach originally proposed for image annotation that learns a joint model of text and image features (**WSABIE**) [20], (2) linear regression and SVM models that use the concatenation of the topic, the caption and the image vectors as input, **LR (Topic+Caption+VGG)** and **SVM (Topic+Caption+VGG)** respectively. Finally, we test two versions of our own DNN using only either the caption (**DNN (Topic+Caption)**) or the visual information of the image (**DNN (Topic+VGG)**).

Table 1 shows the Top-1 average and nDCG scores obtained. First, we observe that the DNN methods perform better for both the evaluation metrics compared to the baseline methods. They achieve a Top-1 average rating between 1.94 and 2.12 better than the Global PPR, Local PPR, WSABIE, LR and SVM baselines. Specifically, the DNN (Topic+Caption+VGG) method significantly outperforms these models (paired t-test, $p < 0.01$). This demonstrates that our simple DNN model captures high-level associations between topics and images. We should

**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 | - | - | - |

also highlight that the network has not seen either the topic or the image during training which is important for a generic model. In the WSABIE model, linear mappings are learned between the text and visual features. This restricts their effectiveness to capture non-linear similarities between the two modalities.

The DNN (Topic+Caption) model that uses only textual information, obtains a Top-1 Average performance of 1.94. Incorporating visual information (VGG) improves it to 2.12 (DNN (Topic+Caption+VGG)). An interesting finding is that using only the visual information (DNN (Topic+VGG)) achieves better results (2.04) compared to using only text. This demonstrates that images contain less noisy information compared to their captions for this particular task.

The DNN models also provide a better ranking for the image candidates. The nDCG scores for the majority of the DNN methods are higher than the other methods. DNN (Topic+Caption+VGG) consistently obtains the best nDCG scores, 0.79, 0.80 and 0.81 respectively. Figure 1 shows two topics and the top-3 images selected by the DNN (Topic+Caption+VGG) model from the candidate set. The labels selected for the topic #288 are all very relevant to a *Surgical operation*. On the other hand, the images selected for topic #99 are irrelevant to *Wedding photography*. For this topic the candidate set of labels do not contain any relevant images.

Topic #288: surgery, body, medical, medicine, surgical, blood, organ, transplant, health, patient

(a) *3.0* (b) *2.8* (c) *2.9*

Topic #99: wedding, camera, bride, photographer, rachel, lens, sarah, couple, guest, shot


(d) *0.4* (e) *0.8* (f) *0.8*

**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.

## 5 Conclusion

We presented a deep neural network that jointly models textual and visual information for the task of topic labeling with images. Our model is generic and works for any unseen pair of topic and image. Our evaluation results show that our

proposed approach significantly outperforms the state-of-the-art method [3] and a relevant method originally utilized for image annotation [20].

## References


1. Aletras, N., Baldwin, T., Lau, J.H., Stevenson, M.: Representing topics labels for exploring digital libraries. In: JCDL (2014)
2. Aletras, N., Baldwin, T., Lau, J.H., Stevenson, M.: Evaluating topic representations for exploring document collections. JASIST (2015)
3. Aletras, N., Stevenson, M.: Representing topics using images. In: NAACL-HLT, pp. 158–167 (2013)
4. Aletras, N., Stevenson, M.: Labelling topics using unsupervised graph-based methods. In: ACL (2014)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003)
6. Chaney, A.J.B., Blei, D.M.: Visualizing topic models. In: ICWSM, pp. 419–422 (2012)
7. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: ICML, pp. 160–167 (2008)
8. Croft, B.W., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley, Boston (2009)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)
11. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: ACL-HLT, pp. 1536–1545 (2011)
12. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: ACL, pp. 302–308 (2014)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Smith, A., Hawes, T., Myers, M.: Hierarchy: visualization for hierarchical topic models. In: Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 71–78 (2014)
15. Snyder, J., Knowles, R., Dredze, M., Gormley, M., Wolfe, T.: Topic models and metadata for visualizing text corpora. In: NAACL-HLT, pp. 5–9 (2013)
16. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: ICML, pp. 129–136 (2011)
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**(1), 1929–1958 (2014)
18. Tieleman, T., Hinton, G.: Lecture 6.5-RMSProp. COURSERA: Neural networks for machine learning (2012)
19. Wan, X., Wang, T.: Automatic labeling of topic models using text summaries. In: ACL, pp. 2297–2305 (2016)
20. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Mach. Learn. **81**(1), 21–35 (2010)