Mathematical Problems of Computer Science 45, 138–142, 2016.

Image Visual Similarity Based on High Level Features of Convolutional Neural Networks

Aghasi S. Poghosyan, Hakob G. Sarukhanyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: agasy18@gmail.com, hakop@ipia.sci.am

Abstract

Nowadays, the task of similar content retrieval is one of the central topics of interest in academic and industrial worlds. There are numerous techniques that are both dealing good with structured data and unstructured such as texts, respectively. However, in this paper we present a technique for retrieval of similar image content. We embed images to N dimensional feature space using convolutional neural networks and perform the nearest neighbor search afterwards. At the end, several distance metrics and their influence on the outcome are discussed. We are rather interested in the proportion of related content than in the additional ranking. Thus, the evaluation of results is based on precision and recall. We have selected 6 major categories from ImageNet dataset to assess the performance.

Keywords: Image retrieval, Convolutional neural networks, Distance metrics.

1. Introduction

The Content-Based Image Retrieval(CBIR) [1] has a broad application from professional database searching to search engines in the Internet. In these applications one uses a sample image as a query item and the CBIR system responds with the list of relevant items. Usually, the resulting list uses underlying several factors. First, it uses a method to derive a knowledge about the contents of the image, such as interest point-based detection methods (SIFT [2] SURF [3]). Secondly, it uses extracted features and a distance metric to measure similarity between the query image and the images in the database. At last, one can derive a ranking method in order to show the query result.
We perform the steps desrcibed above using the following procedure. In order to embed the image into the N dimensional feature space we use the feature vector extracted from pool5/7x7_s1 layer of GoogleNet [5] convolutional neural network (CNN). GoogleNet is trained on the ImageNet [4] dataset and highly effective in object classification and localization task. The pool5/7x7_s1 layer is the last layer in the CNN on which the classification task is performed. In other words, it consists of high level features of the image. Later we use and compare the performance of twelve distance measures to assess the similarity. We do not use any additional ranking methods for sorting the result. In the current case, results are sorted according to the similarity between the query image and the images in the database.

The rest of the paper is organized as follows. In the second section we discuss the feature layer of GoogleNet we used. Later, we give an overview to the twelve distance metrics we mentioned above. Afterwards, we show our evaluation results based on precision/recall metrics. At the end we discuss advantages and disadvantages of this method and further improvements.

2. High Level Feature Extraction

The GoogleNet is a CNN that aims to localize and classify objects in images. It is trained on the ImageNet dataset and shows 6.6% top-5 accuracy error on image classification task. In GoogleNet the last layer before the classification layer is 1024-dimensional vector and called pool5/7x7_s1. This layer extracts the highest level features of the image and gives the most descriptive information about the objects within it. Thus, the features extracted by pool5/7x7_s1 cover the majority information contained within the image. We structure our content by propagating each image through GoogleNet and storing pool5/7x7_s1 layer information to feature vector database (FVD). From now on by database we mean FVD. For query images we follow the same procedure, except the saving part. Specifically, we propagate the query image through GoolgeNet to embed the extracted feature vector to the same 1024-dimensional space.

3. Distance Metrics

The next important step for our task is to choose a distance metric that will maximize the performance. Now we briefly discuss each metric under our consideration below. For a pair of N-dimensional real-valued vectors u and v the distances are defined as follows:

$$\sum_i |u_i - v_i|, \quad (Manhattan) \tag{1}$$

$$||u - v||_2, \quad (Euclidean) \tag{2}$$

$$1 - \frac{u \cdot v}{||u||_2||v||_2}, \quad (Cosine) \tag{3}$$

where $u \cdot v$ is the dot product of u and v.

$$\sum |u_i - v_i| / \sum |u_i + v_i|, \quad (Bray - Curtis) \tag{4}$$

$$d(u, v) = \sum_i \frac{|u_i - v_i|}{|u_i| + |v_i|}, \quad (Canberra) \tag{5}$$

where u_i and v_i are 0 for given i, then the fraction $0/0 = 0$ is used in the calculation.

$$\max_i |u_i - v_i|, \quad (Chebyshev) \tag{6}$$

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{||(u - \bar{u})||_2||v - \bar{v}||_2}, \quad (Correlation) \tag{7}$$

where $\bar{u}$ is the mean of the elements of u and x.

We use a subset Image-Net [4] dataset to assess the retrieval performance. We extract 6 major categories in total size of 5500 images. In Fig. 1 the frequency distribution per category is depicted. In order to increase the precision of our measurements we limit each category to have 440 images. In this way we can ensure that recall metric will not depend on the sample size of the category and will provide more reliable results.

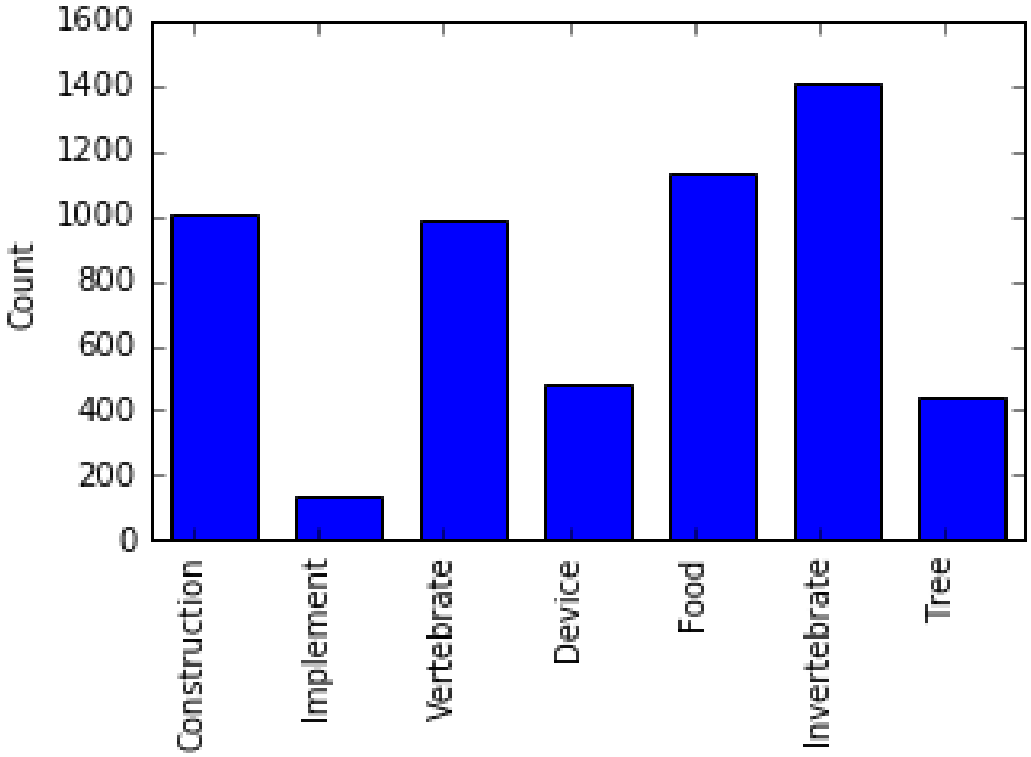

Fig. 1. Image-Net subset per categories image count bar.

4. Evaluation

Precision and recall are common measures on evaluating performance of information retrieval tasks. Precision is defined as the percentage of relevant items in retrieved items. And recall as the percentage of relevant items that were retrieved in total relevant items. Specifically

$$\text{precision} = \frac{|\{\text{relevant items}\} \cap \{\text{retrieved items}\}|}{|\{\text{retrieved items}\}|} \tag{8}$$

$$\text{recall} = \frac{|\{\text{relevant items}\} \cap \{\text{retrieved items}\}|}{|\{\text{relevant items}\}|} \tag{9}$$

In our experiment we fixed then the number of retrieved items (neighbors) to 100. Per image category average precision has been computed by:

$$P_q{}' = \sum_{k \epsilon A_q} \frac{P(i_k)}{|A_q|}, q = 1, 2, ...N, \tag{10}$$

where N is the number of categories, A_q is qth category, $P(i_k)$ is the precision for k th image.

Per image category average recall has been computed by:

$$R_q{}' = \sum_{k \epsilon A_q} \frac{R(i_k)}{|A_q|}, q = 1, 2, ...N, \tag{11}$$

where N is the number of categories, A_q is qth category, $R(i_k)$ is recall for k th image.

The average precision and average recall are given by:

$$P' = \sum_{q=1}^{N} \frac{P_q{}'}{N}, \tag{12}$$

and

$$R' = \sum_{q=1}^{N} \frac{R_q{}'}{N}. \tag{13}$$

In order to evaluate our method we constructed a similarity matrix for the image database. For each image within the image database we retrieved 100 neighbors and calculated precision and recall for each of them. First we have evaluated results per distance metric and averaged them across image categories: Tables 1, 2. We can see that on average correlation metric performs better than the other metrics. However, there are image categories where correlation shows lower results than the other metrics.

Table 1. Metrics comparison precisions.

	Manhattan	Cosine	Euclidean	Braycurtis	Canberra	Chebyshev	Correlation
Construction	0.862	0.928	0.852	0.924	0.912	0.635	**0.932**
Vertebrate	0.603	0.781	0.561	0.782	**0.813**	0.440	0.800
Device	0.814	0.835	0.768	0.836	**0.856**	0.429	0.845
Food	**0.974**	0.949	0.963	0.945	0.948	0.717	0.956
Invertebrate	0.699	0.851	0.654	0.839	0.838	0.364	**0.865**
Tree	0.953	0.944	**0.960**	0.944	0.917	0.795	0.940
Mean	0.818	0.881	0.793	0.878	0.881	0.563	**0.890**

Table 2. Metrics comparison recall.

	Manhattan	Cosine	Euclidean	Braycurtis	Canberra	Chebyshev	Correlation
Construction	0.196	0.211	0.193	0.210	0.207	0.144	**0.212**
Vertebrate	0.137	0.177	0.127	0.177	**0.184**	0.100	0.181
Device	0.185	0.189	0.174	0.190	**0.194**	0.097	0.192
Food	**0.221**	0.215	0.219	0.214	0.215	0.163	0.217
Invertebrate	0.159	0.193	0.148	0.190	0.190	0.082	**0.196**
Tree	0.216	0.214	**0.218**	0.214	0.208	0.180	0.213
Mean	0.185	0.200	0.180	0.199	0.200	0.128	**0.202**

5. Conclusion

This paper investigated another technique for CBIR and evaluated the proposed method against different similarity measures. We have shown that using high level features from GoogleNet in combination with correlation distance metric can lead to promising results.

References

[1] R. Datta, J. Li and J. Z. Wang, "Content-based image retrieval - approaches and trends of the new age", *The Pennsylvania State University, University Park, PA 16802*, 2005.
[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[3] R. Funayama, H. Yanagihara, L. Van Gool, T. Tuytelaars and H. Bay, "Robust Interest Point Detector and Descriptor", *US Patent office*, no. 8165401, 2009.
[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", *IEEE Computer Vision and Pattern Recognition*, 2009.
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions", *arXiv*, no. 1409.4842, 2014.

Submitted 15.09.2015, accepted 25.01.2016

Պատկերների տեսողական նմանությունը հիմնված փաթույթային նեյրոնային ցանցերի բարձր կարգի հատկությունների վրա

Ա. Պողոսյան, Հ. Սարուխանյան

Ամփոփում

Աշխատանքում ներկայացված է փաթույթային նեյրոնային ցանցերի՝ բարձր կարգի հատկությունների վրա հիմնված տեսողական նման պատկերների որոնման համակարգի վերլուծություն: Ցույց է տրվում, որ GoogleNet-ի բարձր կարգի հատկությունների վեկտորների համար հեռավորության ֆունկցիայի ընտրությունը մեծ ազդեցություն ունի որոնող համակարգի ճշգրտության վրա, և վերջինիս համար լավագույն արդյունքները ստացվում են, երբ կոռելյացիան դիտարկվում է որպես հեռավորության ֆունկցիա:

Визуальная схожесть изображений, основанная на свойствах высшего уровня конволюционных нейронных сетей

А. Погосян, А. Саруханян

Аннотация

В данной работе представлено исследование системы поиска визуально похожих изображений, основанной на свойствах высшего уровня конволюционных нейронных сетей. Показано, что выбор функции расстояния для свойств высшего уровня GoogleNet имеет большое влияние на точность поисковой системы. Для получения лучших результатов из просмотренных функций выделяется функция корреляции.