International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:12, No:7, 2018

Improving Similarity Search Using Clustered Data

Deokho Kim, Wonwoo Lee, Jaewoong Lee, Teresa Ng, Gun-Ill Lee, Jiwon Jeong

***Abstract*—**This paper presents a method for improving object search accuracy using a deep learning model. A major limitation to provide accurate similarity with deep learning is the requirement of huge amount of data for training pairwise similarity scores (metrics), which is impractical to collect. Thus, similarity scores are usually trained with a relatively small dataset, which comes from a different domain, causing limited accuracy on measuring similarity. For this reason, this paper proposes a deep learning model that can be trained with a significantly small amount of data, a clustered data which of each cluster contains a set of visually similar images. In order to measure similarity distance with the proposed method, visual features of two images are extracted from intermediate layers of a convolutional neural network with various pooling methods, and the network is trained with pairwise similarity scores which is defined zero for images in identical cluster. The proposed method outperforms the state-of-the-art object similarity scoring techniques on evaluation for finding exact items. The proposed method achieves 86.5% of accuracy compared to the accuracy of the state-of-the-art technique, which is 59.9%. That is, an exact item can be found among four retrieved images with an accuracy of 86.5%, and the rest can possibly be similar products more than the accuracy. Therefore, the proposed method can greatly reduce the amount of training data with an order of magnitude as well as providing a reliable similarity metric.

***Keywords*—**Visual search, deep learning, convolutional neural network, machine learning

I. Introduction

DEEP learning using Convolutional Neural Network (CNN) has been widely known as a generalized solution to solve problems such as image classification, localization, etc. [1]. Answers of these problems are very simple and concise, but they demand very high-level of inferences to solve these problems. For example, “Is this a cat or dog?” or “Where is the cat in the image?” On the other hand, problems resulting in complex answers are still remained and unsolved. Measuring similarity is one of the problems and has become important as demands for the visual search technique from online-markets rapidly increase due to the product recommendation systems. Thus, the visual search techniques have been actively developed by major online-market providers such as Amazon, Baidu, and Taobao. Furthermore, the techniques have also been included in mobile platforms such as Bixby Vision and Google Photos in order to improve user experience as an intelligent system.

Previous studies for the visual search have proposed two main types of training methods in order to make train similarity with deep learning: supervised or unsupervised. The supervised learnings require ground-truth data, called metric data, to train similarity in a quantitative way [2]. The unsupervised learnings, on the other hand, train the model indirectly without any quantitative similarity data, which usually used data for classification [3].

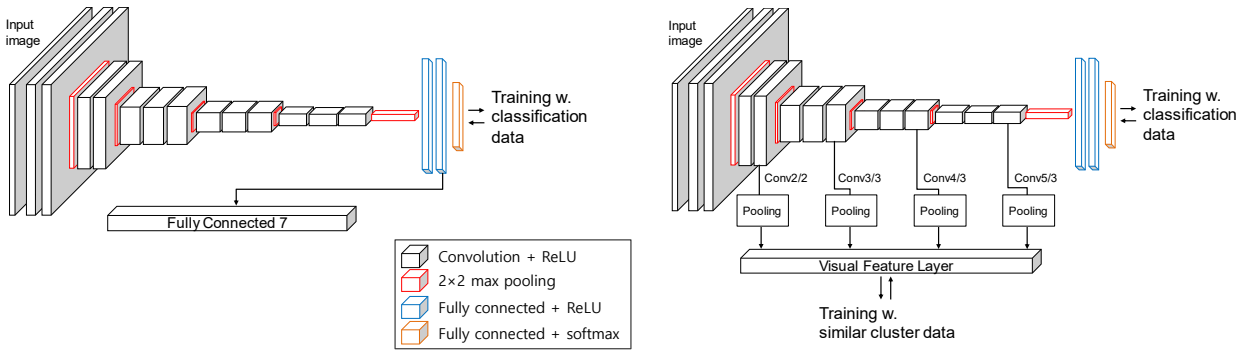
Trivially, the two learning methods have trade-offs between accuracy and efficiency. The supervised learnings can provide accurate results, however, require an extremely large amount of data with size of an order of N^2 where N is the number of images. In practice, gathering the metric data are an unachievable goal when the number of objects to be identified increases. On the other hand, the unsupervised learnings require a relatively small amount of data. The methods train models using data from a different domain. Then, extractions of intermediate data from the trained model are used as visual features to compare similarity distance. These methods are more practical since the size of data for classification has an order of N. However, accuracy of the similarity measurement comes with the unsupervised learnings may not resemble human perception, and cannot be enhanced once the dataset is decided.

In order to overcome limitations of the two learning methods, this paper proposes a *cluster based similarity learning method*. The proposed method is a supervised learning method, but it requires ground-truth data with size of an order of N, which is practical in use. Two non-metric datasets are used for training: a classification datum and a cluster datum. The cluster data contain sets of images, and each set called cluster which consists of images with identical or similar objects. Therefore, the proposed method is able to accurately retrieve similar objects using non-metric data, improving 26.6% of Top-4 search accuracy compared to the conventional method. In addition, the proposed method also achieves 3.2% of Top-1 classification accuracy.

II. Learning Similarity from Non-Metric Data

Comparing similarity is a domain-specific problem since the basis of the similarity comparison varies according to the pair of images. For example, human perception focuses more on the texture for packaged products and the shape for bags. Thus, a network should work differently according to the domain of interests. In order to achieve this, we first train a CNN model as a classification model for target domains. Then, the model is trained by the cluster data. Therefore, the proposed method can produce the classification result and the visual feature with a single network, simultaneously.

D. Kim, W. Lee, J. Lee, T. Ng, G.-I. Lee and J. are with Samsung Research, Samsung Electronics, Seoul, Korea (email: deokho16.kim@samsung.com, wonw.lee@samsung.com, jw84.lee@samsung.com, nkk.teresa@samsung.com, gunill.lee@samsung.com, jiwon.jeong@samsung.com).



(a) (b)

Fig. 1 (a) Conventional unsupervised learning visual feature (deep feature) model, (b) Proposed cluster based leaning model

A. Classification

The proposed method arranges the classes into two levels. The first level named category presents a set of classes, that we want to identify, and the second level presents fine-grained classes. Each category can represent different domains such as clothes, packages, etc. In order to enhance domain-specific characteristics the proposed method, we train the VGG-16 model [3] with a two-step fine-tuning method.

At the first step, low level layers, which are the first four layers near input, are not trained. Then, all the layers are trained at the second step. Thus, the lower layers are less sensitive to train loss for classification; as a results, it preserves the diversity of the pre-trained model's low level characteristics, and it also fine-tunes the low level layers to have domain-specific characteristics.

Preserving diversity of low level layers is important. In fact, the amount of domain-specific classification data is usually small compared to ImageNet data [4], and training with such a limited data can result in convergence of intermediate data in networks which are irrelevant to identifying the classes. Color data is one example, which is converged, when the network tries to identify objects only by shape. In addition, evaluation results of the proposed classification method showed that confusions between classes belong to different categories are reduced. Therefore, networks trained by the proposed method can provide diversified low level data with high accuracy of classification results.

B. Training Visual Feature

Previous state-of-the-art unsupervised learning method uses a single latent layer (the last fully connected layer: FC7) as a visual feature [3]. In opposed to that, the proposed method uses a new layer which is linked to multiple latent layers as a source of visual feature, and the layer is trained with clustered data. The two different methods are depicted in Fig. 1. In fact, the data characteristics of latent layers become high-level of abstraction as passing through the layers. Thus, selecting multiple latent layers can provide diverse characteristics, and we have preserved diversity of low level data using two step training method for classification. Among the layers in the VGG-16 network, the last convolution layers for each group of layers are chosen as the latent layers except the layers in the first group. Thus, layers of 4, 7, 10, 13 are selected and are depicted as Conv2/2, Conv3/3, Conv4/3, and Conv5/3 in Fig. 1 (b), respectively.

In order to reduce dimensionality of feature map data generated from the latent layers, we have tested various pooling methods. As seen in Fig. 2, the spatial region of a feature map can be separated into two different types: global and spatial. Global pooling reduces spatial dimension to one, on the other hand, spatial pooling generates $m \times n$ of $1 \times 1 \times D$ pooled data for regions using moving window. Average or maximum value is chosen for each spatial region. Thus, each pooled feature map becomes a single vector, and four different pooling methods were used by jointly combining global/spatial and average/max pooling.

Then, the pooled data are serialized and connected to a fully connected layer, named visual feature layer in this paper, with output length of 128, and thus the visual feature becomes a vector of 128 elements. By using different pooling methods, three different deep learning models are proposed: 1) GA, 2) SAP, and 3) SAM. GA uses global average pooling for all the selected latent layers. SAP uses spatial average pooling (32×32 window of stride 16) for all the selected layers. SAM uses spatial average pooling (32×32 window of stride 16) for Conv2/2 and Conv5/3 and spatial max pooling (32×32 window of stride 16) for Conv3/3 and Conv4/3.

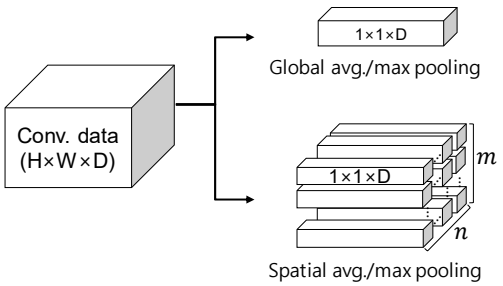


Fig. 2 Feature map pooling methods

Spatial pooling can preserve spatial information regardless of the size of feature map from latent layers. Using max pooling for middle layers can enhance signals related to the objects. Indeed, signals in Conv2/2 still contain data related to

background. On the other hand, signals reached to Conv5/3 loose most of the spatial information. Therefore, middle layers include relatively balanced data related to object with spatial information, which is the reason why max pooling is used for middle layers.

In order to train the visual feature layer with cluster data, triplets are fed to the network, which consist of two images in identical cluster and one image from the other cluster. A hinge loss with cosine similarity distance is used as training loss. Thus, training the networks minimizes the following loss value:

$$Loss(d_+, d_-) = \begin{cases} \max(0, d_+ - 0.5) \\ \max(0, 0.5 - d_-) \end{cases}$$

where

$$d(u, v) = 1 - \frac{u \cdot v}{\|u\|\|v\|}$$

and d_+ and d_- represent the distances of image pairs from an identical cluster and distance of image pair from different cluster, respectively.

The feature is trained to minimize the distances within a cluster and to maximize distances among different clusters.

III. Related Works

Traditionally, machine learning using Neural Networks (NN) had been widely known as a generalized solution to achieve complex, high-dimensional, and nonlinear mappings. However, increasing the size of networks subsequently increases the amount of training data and the learning time. Thus, NN had not been focused for an effective solution until CNN became popular.

Lecun et al. first introduced CNN for handwriting recognition named LeNet-5 [5]. The CNN consists of multiple convolution filters, followed by fully connected layers which are layers of traditional NN. The convolutional layer contains much fewer number of parameters compared to fully connected layers, and the parameters are shared spatially. Therefore, CNN can reduce the number of parameters as well as providing position invariance characteristics due to the shared parameters.

Many years later, Krizhevsky et al. introduced a large CNN network named AlexNet [6]. AlexNet succeeds the structure of LeNet-5. However, the network can process a large size of input image (224×224), and can accelerate the training using multiple GPUs. The generalized classification CNN outperformed previous classification methods by 9.4% of Top-5 error and won the competition, ILSVRC-2012.

Following the success of AlexNet, depth of CNN has been rapidly increased. Simonyan and Zisserman introduced the VGG network [3], which won ILSVRC-2014. They analyze performance variation in terms of depths of a CNN network, which results twice deeper network than AlexNet.

A team from Google introduced their first deep learning network GoogleNet with 22 layers [7], which is deeper than the VGG networks and it also diversifies the network path, called Inception module. In the Inception module, the 1×1 convolutions were used after general convolutional filters, and thus the network achieves a deeper network with small number of operations. From then, Google has revised the Inception module by factorizing the convolutional filters, and now Inception-v4 has been introduced.

A team from Microsoft introduced the deepest CNN network architecture, which has 152 layers [8]. They proposed a residual path, which adds identity data from the previous layer. With the help of the residual path, CNN can have extremely large number of layers without failure of training.

While the classification networks are going deeper with improvement, networks for similarity search have only a few advances. Among them, the Siamese network architecture is the most generalized method in order to train similarity between objects [9], which guides the network to learn metric between multiple images which consist of usually a pair or a triplet of images. As the network directly trains metric, such a supervised learning method can provide the best accuracy on similarity measure. However, the method requires an extremely large amount of data, and thus only a limited domain can utilize the method such as face recognition, identification of facial representation [9], or similar bird identification [10].

In order to avoid the limitation of supervised learning, unsupervised learning methods are widely used in similar image retrieval. The methods train the CNN networks with data from different domain, which is usually classification data, and then use internal data of networks as a feature vectors [3] [11]-[14].

Razavian et al. proposed utilizing intermediate data of CNN as a feature vector [13]. They evaluated various spatial samplings of feature map which are the result data of a convolutional filter. The feature map data were sampled from different spatial grids by using average/max pooling. As a result, sampling the feature map data outperformed previous retrieval methods using feature vectors: FV (Fisher Vector), VLAD (Vector of Locally Aggregated Descriptors), and BoW (Bag of Words).

Mohedano et al. proposed BoW using feature map data, which called local CNN features [14]. The proposed method clusters spatially separated feature map data using *k*-means clustering, and transforms the clustered vectors to BoW representation. Therefore, the proposed method shows better performance on retrieving compared to the Razavian's method.

Lin et al. [3] and Cao et al. [11] introduced methods to extract hash representation from CNN networks. Both methods utilize data from the last fully connected layer, and convert activations of the layer to binary representation. The proposed methods showed performance improvements on retrieving similar fashion items.

Huang et al. proposed a similar method for utilizing data from fully connected layers [12]. However, they additionally train the network attribute based data with classification. The dataset consists of 5-9 semantic attribute categories with more than a hundred attributes for the images.

IV. Experimental Results

The proposed methods were evaluated with dataset for product search including packaged products, clothes, etc. The

classification and the cluster dataset consist of 34 classes of 17 categories and 3628 clusters, respectively.

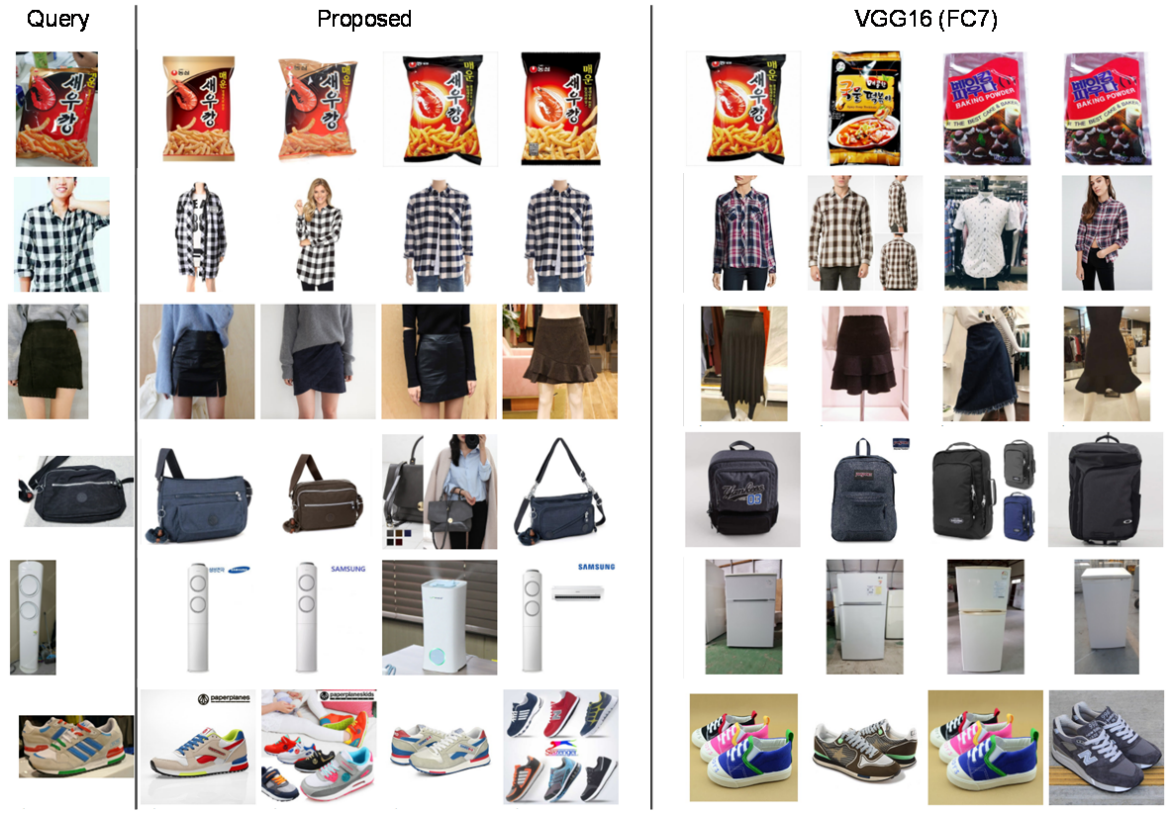


Fig. 3 Evaluating product recommendation performance from 2 million of online market database

Performance improvement of the proposed two-step classification method is presented in Table I, which improves 3.6% and 3.2% of Top-1 classification accuracy for class and category, respectively.

TABLE I
CLASSIFICATION ACCURACY

	Top-1 (class)	Top-1 (category)
Baseline	91.6%	94.8%
Proposed	95.2%	98.0%

Visual search accuracy of the proposed method for image retrieval is presented in Table II. The retrieval accuracy is measured a cluster database with 322 clusters of 644 images. The proposed method was compared with the approach (FC7) introduced by Lin et al. [3]. The accuracy is counted when any image in the identical cluster with the query image is retrieved in Top-4.

As shown in Table II, the performance of the proposed method in image retrieval outperforms the visual feature of previous study (FC7) by 26.6%. Also, using average and max pooling for spatial regions improves 3.1% of accuracy. In fact, global average pooling without spatial region shows better performance compared to the spatial pooling using only average pooling since the separating spatial regions increase the noise data outside object. However, max pooling of middle layers effectively enhances object data, improving the 1.3% of accuracy of 1.3%. In addition, evaluation on real product database was also performed. The database consists of 2 million images from online market providers. Fig. 3 shows the retrieval results of various categories of image queries for the proposed method (SAM) and VGG16 (FC7). As shown in the figure, the proposed method can find significantly better similar images compared to the previous method.

TABLE II
VISUAL SEARCH PERFORMANCE

Model (feature)	Top-4 accuracy
VGG16 (FC7)	59.9%
Proposed (global avg. pool)	85.2%
Proposed (spatial pooling, avg. pool)	83.4%
Proposed(spatial pooling, avg. & max pool)	86.5%

V. CONCLUSION

The proposed method achieved 86.5% of accuracy for finding exact items among Top-4, while the state-of-the-art technique achieved 59.9% of accuracy. Such an improvement can provide one of four retrieved images which is identical to query with a probability of 86.5%, which also means that the rest of retrieved images also shows similar products with a

probability higher than 86.5%. The proposed method can find more similar products compared to the state-of-the-art technique, which is way more similar subjectively. Therefore, the proposed method greatly enhances user experiences on visual search.

The proposed technique trains a deep learning network with cluster data, and therefore, we can reduce the required amount of data significantly, which is 49 thousand times smaller compared to metric data for training similarities between 10 thousand images. In addition, products are usually searched and retrieved in partitioned data by classes to limit the searching time. Performance improvement on classification accuracy can enable partitioning products in database more accurately. Thus, accurate classification can also improve the retrieval results by finding products which have not been partitioned properly.

In conclusion, the proposed method can show remarkable object identification performance with a single model, and the proposed method can continuously improve the performance by collecting cluster data, which is more practical compared to the metric data.

REFERENCES

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, 2015, pp. 85-117.

[2] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," *International Workshop on Similarity-Based Pattern Recognition*. Springer, Cham, 2015.

[3] K. Lin, H. F. Yang, J. H. Hsiao and C. S. Chen, "Deep learning of binary hash codes for fast image retrieval," *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, MA, 2015, pp. 27-35.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015.

[5] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *in Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, 2015, pp. 1-9.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778.

[9] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 539-546, vol. 1

[10] C. Huang, C. C. Loy, and X. Tang, "Local similarity-aware deep feature embedding," *Advances in Neural Information Processing Systems*. 2016.

[11] Z. Cao, M. Long, J. Wang, and P. S. Yu, "Hashnet: Deep learning to hash by continuation*." arXiv preprint arXiv:1702.00758* (2017).

[12] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 1062-1070.

[13] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications,* 4.3, 2016, pp. 251-258.

[14] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i-Nieto, "Bags of local convolutional features for scalable instance search." *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval,* ACM, 2016.