Meta-Webly Supervised Learning for object recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently million of data is shared publicly via Internet. Computer vision researchers have shown interest in learning the object's visual representation through images obtained from the web (WSL). However, it is common that results from search engines return a large number of irrelevant images for the physical environment (context) in which the query was made (e.g. Google Images presents results of *apple fruit*, *apple brand* and *devices* from *apple* query, however, if the query is made at home, *apple fruit* images are more relevant). Recent works in WSL do not consider the context to obtain relevant images of a new object. In this research, we extract meta-learning attributes of objects and their context to identify relevant images of unknown categories, for later using them as training examples in object recognition tasks. Experimental results show that our approach is highly competitive to manually labeled images and to a state of the art curriculum design method for WSL.

1 Introduction

Deep learning models have achieved high performance in object recognition tasks (e.g. Inception V3 [14]), these models have been trained using large-scale manually labeled datasets (e.g. ImageNet[3]), in a fully-supervised approach. Object recognition is part of the everyday tasks of a domestic assistance robot [7],[11],[12],[6], where sometimes it is necessary to identify a non-available object in these datasets. To recognize a new object, we could retrieve images from the Internet (e.g. Google Images), and then manually select the relevant ones, which is expensive cost and time-consuming. Thus the growing interest in Webly Supervised Learning (WSL) for object recognition, which consists in querying the web, download and filter images, and then train a classifier. Recent works in WSL focus in building models or techniques that learn the object's visual representation diminishing the impact of irrelevant images (noise).

2 Proposed method and experimental results

To learn an object's visual representation directly from the web produces a low performance in the object recognition task due to a large amount of noise. Usually researches such as [2], [4], [10], [5] use visual information to diminish the noise's impact. In contrast, we use multi-modal (visual and textual) information to filter images via meta-learning.

Meta-learning allows to learn the images filtering task based on previous experiences (other categories) through a set of descriptive features (meta-features). Our method, receives the object and context as query, it then extracts meta-features based on visual and textual (object and context depending) queries from the images and textual meta-data (title, description, website, subtitles and text from the web page) downloaded. Next we train a classifier using leave-one-out validation to label images as relevant or irrelevant and then unknown object images, labeled as relevant, are used as training set for the object recognition task.

To extract textual meta-features dependent on context and object, we obtain additional information through ConceptNet[13] such as object and context actions, properties and parts, considering that each relevant image's meta-data contains similar words to the queries. Similarity between each query and meta-data is measured using euclidean distance, sum of squared differences, sum of absolute differences, cosine similarity, correlation coefficient with the average vector obtained from Word2Vec[9] and Word Mover's Distance (WMD)[8], using them as meta-features. Similar to [1],[4],[10] we take advantage from top images. Experimentally we determinate to use the average feature vector obtained from Inception V3 of the top 75 images as visual query. The same applied measures in text are used as meta-features, except for WMD which was designed only for text.

We measure the method performance in images filtering task for unknown categories over a list of 27 objects commonly found at home[6]. In this phase, Linear Discriminant Analysis (LDA) archived better results than other classifiers. We tested different ways to join the visual and textual representation such as, early fusion of *context+visual* (which concatenates visual and context meta-features), late fusion, with soft voting, hard voting and stacking (adding logistic regression for final prediction). To measure the impact of the proposed method in recognition task, we feed Inception V3 during the training phase with the filtered images. As baseline we consider the trained model with all the downloaded images. Tables 1 and 2 show that using the visual representation as filtered method is the nearest to the manually labeled images and that all proposed methods improve the baseline and the method presented in [5] in the object recognition task.

Table 1: Results by LDA on image filtering task with single representation, early and late fusion.

Fusion	Information	Precision	Recall	F1-Measure
Single	Object	70.47±26.48%	73.74±27.92%	65.96±26.92%
	Context	68.35±28%	78.17±16.76%	67.68±20.38%
	Visual	**78.98±25.88%**	83.15±14.47%	**79.12±20.89%**
Early	Object+Context	69.95±26.26%	73.24±27.19%	65.77±25.86%
	Object+Visual	77.70±26.4%	79.80±21.9%	75.65±23.25%
	Context+Visual	*78.85±25.48%*	82.68±14.04%	*78.47±20.2%*
	Object+Context+Visual	77.74±26.19%	79.38±22.59%	75.05±23.65%
Late	Soft voting	74.37±26.07%	**83.82±16.81%**	75.8±21.65%
	Hard voting	72.18±27.09%	81.76±18.09%	72.99±22.36%
	Stacking	78.20±26.05%	*83.17±16.7%*	78.43±21.44%

Table 2: Results in object recognition task with Inception V3.

Filtered method	Precision	Recall	F1-Measure
Manually Labeled	78.30±12.43%	87.11±11.49%	81.85±9.95%
From Google Images	65.57±15.68%	80.89±20.46%	71.55±16.23%
Visual	**71.75±15.93%**	*82.22±16.91%*	**75.90±14.91%**
Context+Visual	*71.71±15.67%*	**82.37±16.38%**	*75.88±14.45%*
Soft voting	71.14±17.31%	81.04±19.52%	74.84±16.67%
[5]	64.78±16.28%	73.19±18.23%	68.89±15.6%

3 Conclusions

We present a method based on meta-learning with multi-modal information capable of selecting relevant images from unknown categories via meta-features. Our method is highly competitive with manually labeled images, it obtains better results than one a state of the art method[5] for WSL and represents a good alternative to learn an object's visual representation through web images.

References

[1] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[2] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014.

[5] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[6] R. Izquierdo-Cordova, E. F. Morales, L. E. Sucar, and R. Murrieta-Cid. Searching objects in known environments: Empowering simple heuristic strategies. In *Robot World Cup*, pages 380–391. Springer, 2016.

[7] T. Kollar, M. Samadi, and M. Veloso. Enabling robots to find and fetch objects by querying the web. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 1217–1218. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[8] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] D. Modolo and V. Ferrari. Learning semantic part-based models from google images. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[11] M. Samadi, T. Kollar, and M. Veloso. Using the web to interactively learn to find objects. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[12] M. Samadi, M. Veloso, and M. Blum. Openeval: Web information query evaluation. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[13] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.