AI600 Lab at ImageCLEF 2019 Concept Detection Task

Xinyi Wang[1] and Ningning Liu[2]

[1] School of International Trade and Economics,
[2] School of Information Technology and Management,
University of International Business and Economics, Beijing 100029, P.R.China
iwangxinyi@163.com; ningning.liu@uibe.edu.cn
http://lab.uibe.edu.cn

Abstract. In this paper we describe the participation of AI600 Lab in the ImageCLEF 2019 Concept Detection task. We adopted an approach based on bag-of-visual-words model and logistic regression, using different SIFT descriptors as visual features. The classifiers were trained with different features respectively and weighted results were presented. Our best result ranked 26th among 58 runs and 7th out of 11 participant teams.

Keywords: Concept Detection, Bag of Visual Words, Logistic Regression, ImageCLEF

1 Introduction

In the previous ImageCLEF medical tasks, a lot of remarkable works have been proposed. While traditional methods and features were used [1-3], methods based on deep learning were also introduced [3-4]. In this year, ImageCLEF 2019 [5] Concept Detection task [6] aims on interpreting and summarizing the insight of radiology medical images automatically. For this task, we focused on multi-label classification with traditional visual features.

The remainder of this paper is organized as follows: Section 2 introduces the detailed process of our experiment. Section 3 summarizes all of our submissions. Finally, in Section 4, we make a brief conclusion of our results.

2 Experiments

2.1 Data description

This task used a subset of the Radiology Objects in COntext (ROCO) dataset [7]. Three image datasets were provided. The training, validation and test datasets con-

tained 56,629, 14,157 and 10,000 radiology images. The training and validation sets were accompanied by UMLS concepts extracted from the original image caption. No external data were used in our participation.

The training and validation sets were labeled with a total of 5,528 different concepts. We obtained the frequency distribution of all concepts. The distribution is showed in Table 1. Most of the concepts rarely appeared in the dataset. Only 58 of the 5,528 concepts were labeled with for more than 1000 times. Some *major* concepts appeared frequently in the image set while most concepts were difficult to detect.

Table 1. Frequency statistics of the concepts in training set

Frequency	Number	Proportion
0-10	3718	67.26%
10-100	1261	22.81%
100-1000	491	8.88%
>=1000	58	1.05%
Total	5528	100.00%

Besides, we noticed that many labels are linked and correlated. For instance, images labeled with Concept B in Table 2 were always labeled with Concept A. Among the concepts which were annotated with for more than 100 times in the training set, there were 157 pairs of concepts with strict inclusion relation. This relation was used for detecting some *minor* concepts.

Table 2. Examples of concept pairs of Concept B(subset)-Concept A(superset)

Concept B	Freq.	Concept A	Freq.
C0729233: dsct of thoracic aorta	843	C0817096: thoracics	7470
C3244306: operations	248	C0543467: surgically	1386
C0175676: echotomography	925	C0041618: medical sonography	3257
C0392148: presences	783	C0150312: found	1354
C0203379: 4d echocardiogr	734	C0183129: echocardiographs	1495
……		……	

2.2 Visual features and Bag-of-Visual-Words model

We employed 4 kinds of SIFT descriptors as visual features: SIFT [8], C-SIFT [9], HSV-SIFT [10] and RGB-SIFT. A series of key points of all kinds of descriptors were extracted from each image. To build a bag-of-visual-words (BoVW) model, 2 million key points were randomly selected from the training set as the template key points of visual codebooks. To overcome the memory limitation, we calculated visual codebooks using mini batch k-means [11], a variant of k-means algorithm. Compared to k-means algorithm, mini batch k-means can reduce the amount of computation and work faster. We tried various codebook sizes, or numbers of cluster centroids, and eventually used two different sizes: k = 10,000 and k = 20,000.

For all images, histograms of features were calculated with different codebooks. Each extracted key point in an image was assigned to its closest clustering in the codebook by calculating the Euclidean distance to the cluster centroids. Then the frequency of different clusters was calculated as the representations of images.

Finally, the Term Frequency – Inverse Document Frequency (tf-idf) weights of visual words frequency matrices were calculated and normalized by the L1-norm.

2.3 Classification

We employed a two-round classification. As the distribution of concepts was unbalanced, we dropped most of the concepts and only considered *major* concepts which appeared in the training set more than a frequency threshold, *F*. *F* ranged from 800 to 1,500. After the first stage of classification, the matrices fed into the model were augmented with ground truth or predicted values of the appearances of *major* labels, then some *minor* concepts which were subsets of the concepts predicted and appeared more than 100 times were predicted. This improved the performance of the model slightly.

We applied logistic regression as we deemed it a competitive and faster method of classification compared to support vector machine or k-Nearest Neighbor cluster. For this multi-label classification task, we trained classifiers for each concept separately. Each time we only used one feature for training and prediction. The final submissions were generated from the probabilistic results.

2.4 Experimental environment

Our experiment was conducted under Ubuntu 18.04 operating system with Python 2.7.15. The mini batch k-means clustering and logistic regression algorithm were implemented using scikit-learn library [12]. Some necessary libraries, such as NumPy, Pandas and SciPy were also used. All SIFT visual features were extracted with ColorDescriptor software (version 4.0) [13].

3 Results

3.1 The submitted runs

We submitted 7 runs to ImageCLEF 2019 concept detection task, with 1 run of single feature model and 6 runs of ensemble models. For the ensemble model we weighted the results of single feature model. The weights of [SIFT, C-SIFT, HSV-SIFT, RGB-SIFT] were [0.3, 0.2, 0.2, 0.3]. For the probability threshold *p*, we proposed a method for optimal threshold selection. The probability threshold we used made the concept distribution of the results on test set similar to the concept distribution of best predictions on validation set which had higher F1-scores [14]. We picked a few thresholds in a small range.

The details of submitted runs are as follows.

1. **ai600_result_rgb_1556989393**: single feature model based on RGB-SIFT. The size of visual codebook k = 10,000. The frequency threshold F = 1,200, with 46 *major* concepts and 14 *minor* concepts used for training and prediction. The probability threshold p = 0.1.
2. **ai600_result_weighing_1557059794**: the combination of SIFT, C-SIFT, HSV-SIFT and RGB-SIFT. The size of visual codebook k = 10,000. The frequency threshold F = 1,200, with 46 major concepts and 14 minor concepts used for training and prediction. The probability threshold p = 0.2.
3. **ai600_result_weighing_1557061479**: the same as the ai600_result_weighing_1557059794, except that the size of visual codebook k = 20,000.
4. **ai600_result_weighing_1557062212**: the same as the ai600_result_weighing_1557059794, except that the frequency threshold F = 1,000. In total, 58 *major* concepts, as well as 25 *minor* concepts were used and predicted.
5. **ai600_result_weighing_1557062494**: the same as the ai600_result_weighing_1557059794, except that the probability threshold p = 0.1.
6. **ai600_result_weighing_1557107054**: the same as the ai600_result_weighing_1557059794, except that the frequency threshold F = 1,500. In total, 35 *major* concepts, as well as 8 *minor* concepts were used and predicted.
7. **ai600_result_weighing_1557107838**: the same as the ai600_result_weighing_1557059794, except that the frequency threshold F = 1,000 and the probability threshold p = 0.1. In total, 58 *major* concepts, as well as 25 *minor* concepts were used and predicted.

3.2 Results

The results obtained by our 7 runs are given in Table 3. All 7 runs were graded successfully. The best result of our runs scored a F1-score of 0.1656, which ranked 26th out of 58 runs and 7th out of 11 teams.

Table 3. The results of submitted runs.

Submission Id	Run	F1-Score
27071	ai600_result_rgb_1556989393	0.1345022
27074	ai600_result_weighing_1557059794	0.1628424
27075	ai600_result_weighing_1557061479	**0.1656261**
27076	ai600_result_weighing_1557062212	0.1588862
27077	ai600_result_weighing_1557062494	0.1562828
27095	ai600_result_weighing_1557107054	0.1603341
27096	ai600_result_weighing_1557107838	0.1511505

4 Conclusion

In this paper we have presented the methods we have used in the ImageCLEF 2019 Concept Detection task. We applied multi-label classification based on bag-of-visual-words model with color descriptors and logistic regression. From our experimental results we can conclude the following: (i) while RGB-SIFT descriptors performed best among the color descriptors, the weighted model improved the performance greatly; (ii) using the semantic relations among the concepts, the two-stage classification is able to detect some concepts which are small in number, and on the validation set it can improve the F1-score for about 1%; (iii) with the approach we proposed, it is still challenging to predict concepts with a very limited number of image samples.

References

1. Valavanis, L., Stathopoulos, S.: IPL at ImageCLEF 2017 Concept Detection Task. CLEF working notes, CEUR, 2017.
2. Valavanis, L., Kalamboukis, T.: IPL at ImageCLEF 2018: A kNN-based Concept Detection Approach. CLEF working notes, CEUR, 2018.
3. Pinho, E., Costa, C.: Feature Learning with Adversarial Networks for Concept Detection in Medical Images: UA.PT Bioinformatics at ImageCLEF 2018, CLEF working notes, CEUR, 2018.
4. Wang, X. Zhang, Y., Guo, Z., Li, J.: ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning, CLEF working notes, CEUR, 2018.
5. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Ben Abacha, A., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Cuevas, C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, LNCS Lecture Notes in Computer Science, Springer, 2019.
6. Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Müller, H.: Overview of the ImageCLEFmed 2019 Concept Detection Task, CLEF working notes, CEUR, 2019.
7. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, Proceedings of the MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS 2018), Lecture Notes in Computer Science (LNCS) Volume 11043, pp. 180-189, Springer Verlag, 2018.
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.
9. Burghouts, G.D., Geusebroek, J.M.: Performance evaluation of local color invariants, Computer Vision and Image Understanding, vol. 113, pp. 48–62, 2009.
10. Bosch, A., Zisserman, A., Muoz, X.: Scene classifification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 04, pp. 712–727, 2008.

11. Sculley, D.: Web-scale k-means clustering. Proceedings of the 19th International Conference on World Wide Web. pp. 1177–1178. WWW '10, ACM, New York, NY, USA, 2010.
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, pp. 2825–2830, 2011.
13. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating Color Descriptors for Object and Scene Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 32 (9), pp. 1582-1596, 2010.
14. Liu, N., Dellandréa, E., Chen, L., Trus, A., Zhu, C., Zhang, Y., Bichot, C., Bres, S., Tellez, B.: LIRIS-Imagine at ImageCLEF 2012 Photo Annotation Task. CLEF working notes, CEUR, 2012.