ConceptFusion: A Flexible Scene Classification Framework

Mustafa Ilker Sarac[1], Ahmet Iscen[2], Eren Golge[1], and Pinar Duygulu[1,3]

[1] Department of Computer Engineering,
Bilkent University, Ankara, Turkey
[2] Inria, Rennes, France
[3] Carnegie Mellon University, PA, USA

Abstract. We introduce ConceptFusion, a method that aims high accuracy in categorizing large number of scenes, while keeping the model relatively simpler and efficient for scalability. The proposed method combines the advantages of both low-level representations and high-level semantic categories, and eliminates the distinctions between different levels through the definition of concepts. The proposed framework encodes the perspectives brought through different concepts by considering them in concept groups that are ensembled for the final decision. Experiments carried out on benchmark datasets show the effectiveness of incorporating concepts in different levels with different perspectives.

Keywords: Scene recognition, Concepts, Ensemble of Classifiers.

1 Introduction

With the recent advancements in capturing devices, billions of images have been stored in personal collections and shared in social networks. Due to limitation and subjectivity of the tags, visual categorisation of images is desired to manage huge volume of data.

As an important visual content, scenes have been considered in many studies to retrieve images. Low-level features are commonly used to classify scenes, such as for indoor versus outdoor, or city versus landscape [9, 11, 13–15]. Alternatively, object detector responses have been used as high-level features to represent semantics [8]. While the number of objects could reach to hundreds and thousands with the recent detectors that can be generalised to variety of catagories, the main drawback of object-based approaches is the requirement for manual labeling to train the object models. Moreover, it may be difficult to describe some images through specific objects. Recently, a set of mid-level attributes that are shared between object categories, such as object parts (wheels, legs) or adjectives (round, striped) [3, 6], have been used. However, these methods also heavily depend on training to model human-defined attributes. The main question is how can we melt representations with different characteristics in the same pool? Moreover, how can we scale it to large number of concepts?

In this study, we introduce ConceptFusion, in which we use the term concept for any type of intermediate representation, ranging from visual words to attributes and objects. We handle the variations between different levels of concepts, by putting them into concept groups. Separate classifiers are trained for each concept group. The contributions

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 55–60, 2015.

of each concept group to the final categorization are provided in the form of confidence values that are ensembled for the final decision. The framework is designed to be generalised to large number of different concepts. While early and late fusuion techniques have been studied for a long time, the spirit of our work differs from the others in the following aspects.

- We do not restrict ourselves to only semantic categories that can be described by humans, but also map low-/mid-level representations into concepts.
- Motivated by the recent studies in learning large number of concepts from weakly labeled and noisy web images, the framework is designed to be scaled through the introduction of concept groups.

2 Our Method

ConceptFusion brings the ability of using different levels of descriptors through the definition of concepts and concept groups (see Figure 1). Low-level local or global descriptors could be quantized to obtain concepts in the form of visual words, and then concept group can be represented as Bag-of-Words. On the other hand, each object category could correspond to a concept, and as a whole the concept group could be represented through a vector of confidence values of object detectors. ConceptFusion is designed to allow the integration of different *concept groups* for classification. Concept groups are not required to have any semantic meaning; we suppose that, each concept group can add a different perspective for classification. The classification has two main parts; *individual classification* and *ensemble of classifiers*. Individual classification is applied to each concept group separately, and classification results are combined in ensemble of classifiers stage before making a final prediction.

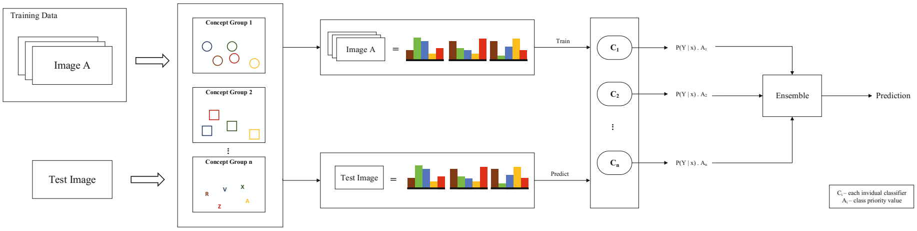

Fig. 1. Overview of ConceptFusion. Individual classifiers are trained for each concept group, and for each individual classifier a concept-priority value is computed. A test image, represented by concepts, is fed to the individual classifiers. Class-confidence values incorporated with the concept-priority values, are combined in the ensemble stage for final prediction.

Individual Classification: To support our hypothesis of trying to examine the different perspectives of each concept group, we consider each group independently. That is, we assume that the individual classification performance of a concept group has no effect on another, and should therefore be treated completely separately. This also allows us to have an agnostic classification method that can be used with any type of *concepts*. To implement this idea, we train a separate, individual probabilistic classifier for each concept group. For a given image query, the role of each individual classifier is to give the probability of the image belonging to each class. We use probabilistic Support Vector Machine (SVM) as classifier.

Ensemble of Classifiers: After training a separate classifier for each concept group, we must be able to combine them properly before making a final decision. Since we cannot guarantee that each individual classifier will perform well, especially in the case of classifiers trained from weakly labeled web images, we decide to explore giving priorities to each individual classifier. To decide which individual classifier gets which priority, we should estimate how a classifier would work on unseen data, so we can assign more weight to decisions of those that are expected perform well, and less weight to those that are predicted to perform poorly.

We introduce the notation of *concept-priority value* as an estimate of how each classifier would perform generally. We find this value by performing cross-validation on the training set using each classifier and assigning the average accuracy value as the *concept-priority value* of the corresponding individual classifier. Now that we have a generalized estimation for the performance of each individual classifier, we can weight their outputs accordingly. Probability outputs of each single classifier is multiplied by its *concept-priority value*. After obtaining the weighted class-confidence probabilities from each classifier, we ensemble them together in the final step. At the end, the class that obtains the highest value is selected as the final prediction.

To demonstrate the ConceptFusion idea, it is desired to include concepts at different levels. To eliminate effort for the manual labeling of objects or attributes, we take the advantage of two benchmark datasets where the semantic categories are already available in some form: MIT Indoor [12] and SUN Attribute Dataset [10].

3 Evaluation of ConceptFusion Framework

In this section, we evaluate ConceptFusion framework to understand the effect of different ensemble techniques, number of concepts and different classifiers.

First, we evaluate the possibilities of using different ensemble methods to combine vectors from different concept groups: *(i) Confidence summation without weighted classifier ensemble* which simply sums the confidence values obtained from classifiers of different concept groups, that is we treat each classifier with equal importance and do not consider any weighting to their results. *(ii) Confidence summation with weighted classifier ensemble* in which before combining the confidence values of each classifier in the summation step, we multiply each of them by the corresponding class priority value. *(iii) Ranking without weighted classifier ensemble*, in which we integrate a classic ranking system [5] to combine different features. Instead of using exact confidence values, we sort the confidence values of each class and rank each class in the order of preference. Then we sum their ranks to come up with a final decision. *(iv) Ranking with weighted classifier ensemble* which weigths the class ranks from classifier by its *concept-priority* value, in order to avoid the possible issues that can rise from treating each classifier equally. *(v) Two-layer classifier as ensemble* where the input of the classifier would be the output of the previous classifiers concatenated together.

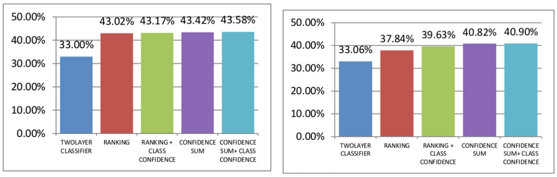

Fig. 2. The effect of ensemble techniques in Sun Attribute (left) and MIT Indoor (right) datasets

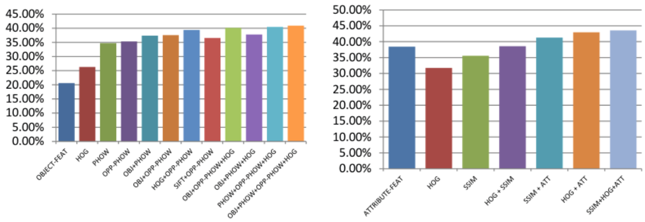

Fig. 3. Comparing different number of concept groups on MIT (left) and SUN (right) datasets

As seen in Figure 2, although changing the ensemble method did not have much effect in the Sun Attribute Dataset, the results of the MIT Indoor Dataset are more distinct. In MIT Indoor, ensembling concept groups using confidence summation and weighted methods is clearly more advantageous than using a ranking system or a non-weighted system. Using confidence- based methods reduces the probability of losing information classifier information, and class-priorities give each classifier their assumed generalized performance rate. We can argue for the same trend in SUN Attribute Dataset, but the difference of accuracies is much less. Two-layer classifiers gives us the worst results for both datasets, because the second level classifier is extremely prone to over-fitting the output of the first layer classifier during the training stage, hence not working well in the testing stage.

Secondly, we evaluated ConceptFusion by changing the number of different concepts used in each dataset. For ensemble of classifiers phase, we use the weighted versions. SVM parameters are set using cross-validation on training data. Results for both datasets are reported in Figure 3. We observe that the accuracy of the classifier also generally increases as we add more concept groups to our system. We obtain the best results by using the highest amount of concept groups. This shows that the combination features from completely different concept groups can be beneficial to the overall classifier, and that our method makes use of this relation in a meaningful way.

Finally, to evaluate the effect of using different classifiers, we used a fixed ensemble configuration and changed the type of our classifier in order to observe any different behaviors. We originally designed ConceptFusion with SVM classifier, however we believe it would also be necessary to see the performance of our framework using two other classifiers: Ada- Boost [4] and Random Forests [1]. As seen on Table 1, LIBSVM's [2] implementation of SVM outperforms the other two classifiers with its capability of constructing non-linear decision boundaries.

Table 1. Comparison of different classifiers on MIT and SUN dataset

	MIT Indoor		SUN Attribute	
	Confidence	Ranking	Confidence	Ranking
Random Forests	37.3%	32.3%	32.7%	33.3%
Ada-Boost	35.8%	33.9%	33.2%	34.7%
SVM	43.6%	43.2%	40.9%	39.6%

4 Comparison with Other Methods

We compare the results of ConceptFusion with a baseline method, and with the state-of-the-art Object Bank method [7] (see Table 2). As the baseline we combine different concepts or features just by concatenating them. This method is extremely simple and widely used, but it can have many disadvantages, such as resulting features being in very high dimensions. Also, combining features from very different concepts, such as low- level and high-level features, does not necessarily add any meaning for classification purposes, and can provide low results. Object Bank [7] is a well known method with the idea of having a higher semantic level description of images, exposing scene's semantic structure similar to human understanding of views. Although ObjectBank provides a good interpretation of the image, it produces a very high dimensional vectors, and concatenation of large number of features does not to perform well.

Table 2. Comparisons with feature concatenation and Object Bank [7] on MIT dataset

Method	Accuracy
Feature Concatenation	9.48%
OB-LR [7]	37.6%
ConceptFusion	40.9%

5 Discussion and Future Work

We proposed ConceptFusion as a framework for combining concept groups from many different levels and perspectives for the purpose of scene categorization. The proposed framework provides flexibility for supporting any type of concept groups, such as those that have semantic meanings like objects and attributes, or low-level features that have no meanings semantically but can provide important information about the structure of an image. There is no limit in the definition of concepts, and it is easy to be expanded through inclusion of any other intermediate representation describing the whole or part of the image in content or semantics.

Current framework examines each concept group on the same level, by assuming that their classification models are completely independent from each other. We plan to extend our framework by modifying this idea, and establishing dependence between each concept group by their semantic meanings.

Acknowledgement. This study was partially supported by TUBITAK project with grant no 112E174 and CHIST-ERA MUCKE project. This paper was also partially supported by the US Department of Defense, U. S. Army Research Office (W911NF-13-1-0277) and by the National Science Foundation under Grant No. IIS-1251187. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ARO, the National Science Foundation or the U.S. Government.

References

[1] Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)

[2] Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2, 27:1–27:27 (2011)

[3] Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)

[4] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)

[5] Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. IEEE PAMI 16(1), 66–75 (1994)

[6] Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by betweenclass attribute transfer. In: CVPR (2009)

[7] Li, L.-J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)

[8] Li, L.-J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classification. In: Kutulakos, K.N. (ed.) ECCV 2010 Workshops, Part I. LNCS, vol. 6553, pp. 57–69. Springer, Heidelberg (2012)

[9] Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)

[10] Patterson, G.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR (2012)

[11] Payne, A., Singh, S.: Indoor vs. outdoor scene classification in digital photographs. Pattern Recogn. 38(10), 1533–1545 (2005)

[12] Quattoni, A., Torralba, A.: Recognizing indoor scenes (2007)

[13] Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T.: A thousand words in a scene. IEEE PAMI 29(9), 1575–1589 (2007)

[14] Serrano, N., Savakis, A.E., Luo, J.: A computationally efficient approach to indoor/outdoor scene classification. In: ICPR (4) (2002)

[15] Vailaya, A., Member, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.-J., Member, S.: Image classification for content-based indexing. IEEE Transactions on Image Processing 10, 117–130 (2001)