Düşük Seviye Görsel Öznitelikler ile Basit Bir Konvolüsyonel Sinir Ağından Elde Edilen Özniteliklerin Birleştirilmesi

Combining Low-Level Image Features with Features from A Simple Convolutional Neural Network

Özge Öztimur Karadağ, Özlem Erdaş
Bilgisayar Mühendisliği Bölümü, Alanya Alaaddin Keykubat Üniversitesi, Antalya, Türkiye
{ozge.karadag, ozlem.erdas}@alanya.edu.tr

***Özetçe*—Görüntü işleme çalışmalarında, geleneksel yaklaşımda öncelikle görüntüden düşük seviye öznitelikler çıkartılır ve daha sonra işlenmek üzere bir tanıma sistemine iletilir. Geleneksel görüntü işleme teknikleri bu adım adım yaklaşımı benimserken, güncel yaklaşımların pek çoğunda hem öznitelikleri çıkartan hem de sınıflandırma veya tanıma işlemini gerçekleştiren katmanlı yapılar tercih edilmektedir. Derin öğrenme teknikleri olarak isimlendirilen bu yapılar yeterli miktarda etiketli verinin mevcut olması ve en düşük sistem gereksinimlerinin karşılanması koşulu ile uygulanabilmektedir. Bununla birlikte, çoğu zaman ya veri miktarı yetersiz olmakta ya da sistem kaynakları karşılanamamaktadır. Bu çalışmada, düşük seviye öznitelikleri basit bir derin öğrenme nöral ağından çıkartılan öznitelikler ile birleştirilerek etkili bir görsel sunum elde etmenin mümkün olduğu deneyimlenmiştir. Sonuç olarak, görüntü veri setimizde birleştirilmiş öznitelikler ile 0.80 doğruluk elde ederken düşük seviye ve derin öğrenme öznitelikleri ile elde edilen doğruluk değerleri sırasıyla 0.70 ve 0.74 olarak bulunmuştur.**

Anahtar Kelimeler—görüntü işleme, öznitelik çıkarma, düşük seviye öznitelikler, konvolusyonel nöral ağlar.

***Abstract*—In the traditional image processing approaches, first low-level image features are extracted and then they are sent to a classifier or a recognizer for further processing. While the traditional image processing techniques employ this step-by-step approach, majority of the recent studies prefer layered architectures which both extract features and do the classification or recognition tasks. These architectures are referred as deep learning techniques and they are applicable if sufficient amount of labeled data is available and the minimum system requirements are met. Nevertheless, most of the time either the data is insufficient or the system sources are not enough. In this study, we experimented how it is still possible to obtain an effective visual representation by combining low-level visual features with features from a simple deep learning model. As a result, combinational features gave rise to 0.80 accuracy on the image data set while the performance of low-level features and deep learning features were 0.70 and 0.74 respectively.**

Keywords—image processing, feature extraction, low-level features, convolutional neural networks.

I. INTRODUCTION

Representation is considered as the initial step of majority tasks in image processing such as segmentation, object recognition, detection etc. This initial step is believed to have a great influence in the system performance. Traditional image processing systems propose various feature extraction schemas and employ the extracted features via a classifier or a recognition/detection system for further processing [1].

Recently, Convolutional Neural Network (CNN) [2] has become very popular in the image processing literature. While the traditional systems first extract features and then give those features to a classification model, CNN processes image data in several layers where it both extracts features and classifies a given image. While the CNN model achieves good performance in classification, both its system requirements and large dataset requirement arise as its major drawbacks. Most of the time a GPU capable device with sufficient amount of memory is required. In order to be able to train a CNN, sufficient amount of labeled data should be available. Alternatively, an already trained network such as AlexNet [3] can be employed. However, it is only possible if the classes of our dataset is already recognized by that CNN. Otherwise, the transfer learning [4], which implies training of an already trained network with new data, can be

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

Figure 1. A sample image for each class from the dataset.

applied as long as enough labeled data is available and the system requirement are satisfied.

This study addresses the problem, where either the system requirements are not sufficient for constructing a complex CNN model, or the number of available labeled data is not enough to train such a model. In that case, CNN model can be employed for feature extraction, and visual representation can be enhanced by combining low-level image features with features from a simple CNN model. In this approach, the representative power of CNN is combined with the simplicity of low-level features.

II. MATERIALS AND METHODS

In this study, two different image representation schemas are investigated and a new representation model which combines the two schemas is proposed. The first representation schema uses low-level image features, while the second one uses the features obtained from a convolutional neural network. In order to compare the two different representation models, features are employed by a classification system and the classification performances are reported for a qualitative comparison.

A. Feature Extraction

1) Low-Level Features

In this study, SIMPLE [5], which employs global descriptors as local ones, is utilized. For this purpose, first Speeded-Up Robust Feature (SURF) [1] detector which employs a Hessian matrix for fast computation and increased accuracy is used to detect regions of interest in an image. SURF is a scale-invariant method since it is robust to orientation and size of images. After applying SURF detector to the data set, Color and Edge Directivity Descriptors (CEDD) [6] are used to extract features from the detected image patches. CEDD features are limited to at most 54 bytes of information even for a large image and combine the color and texture information in a single histogram [6]. The size of the feature vector for a given image is 144, after applying SIMPLE descriptors.

2) Features from CNN

A simple Convolutional Neural Network is constructed for feature extraction. The layers of CNN is arranged as below:

Layer 1- An input layer of size [28 x 28 x 3],
Layer 2 - A convolution layer consisting of 20 mask each with size 3x3,
Layer 3 - A rectified linear unit layer (ReLU),
Layer 4 - A maxpooling layer,

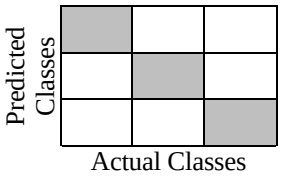

Figure 2. Confusion Matrix

Layer 5 - A fully-connected layer,
Layer 6 - A softmax layer
Layer 7 - A classification layer.

Stochastic gradient descent with momentum method is used for training with learning rate initially set as 0.0001 and the maximum epoch number is set as 10. The output of the fifth layer, which is fully-connected layer, is employed as image features. The size of the feature vector for a given image is 3.

3) Combination of Low-Level and CNN features

For a given image, the combination of low-level and CNN features is obtained by concatenation of low level image features of size 144 with features from the CNN whose size is 3. Hence, the dimension of the combined feature vector for a given image is 147.

B. Classification

Support Vector Machines with linear kernel are employed for the multi-class classification problem.

C. Dataset

A set of images belonging to classes; airplane, ketch and helicopter, from the Caltech 101 [7] dataset is used in our experiments. Sample images from the dataset are provided in Figure 1. Seventy percentage of images are randomly selected for training and the remaining are used for testing.

D. Performance Measure

The classification performances for all three representations schemas are compared using the F-Score criterion. In order to estimate F-score for a classification of the test set, first, the confusion matrix is obtained. For a given set of images and their predicted labels, the confusion matrix depicts the ratio of correct and misclassified images as shown in Figure 2. In this figure, entries of the confusion matrix for a three class classification problem are represented. The number of correct classifications for each class are represented in the diagonals.

Classification performance is commonly evaluated using the precision and recall criteria. Precision is the measure of exactness while recall is the measure of completeness. Using a confusion matrix M, precision and recall are estimated with equation 1 and 2 respectively, and corresponding F-score is estimated with equation 3.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

0.77	0.12	0.12
0.19	**0.70**	0.12
0.27	0.08	**0.65**

Table I. Confusion Matrix for Classification with Low-Level Features

0.92	0.04	0.04
0.15	**0.65**	0.19
0.04	0.12	**0.85**

Table II. Confusion Matrix for Classification with Features from CNN

0.92	0.04	0.04
0.12	**0.77**	0.12
0.08	0.12	**0.81**

Table III. Confusion Matrix for Classification with Combined Features

Features used in classification	Average of Mean F-score
Low-Level Features	0.70
Features from CNN	0.74
Combined Features	0.80

Table IV. Average of Mean F-score values over 10 folds

$$Precision_i = \sum_j \frac{M_{ii}}{M_{ji}} \quad (1)$$

$$Recall_i = \sum_j \frac{M_{ii}}{M_i} \quad (2)$$

$$FScore = 2\frac{Precision*Recall}{Precision+Recall} \quad (3)$$

E. Experimental Setup

The experiments are implemented on MATLAB R2017a on an Intel Core i5-6200U CPU 2.30Ghz. No GPU is employed in our experiments and the time required for the system to extract the features and complete the classification for three different features for 10 folds and then estimate the performances is 153.33 seconds.

F. Results

Confusion matrix for classification results using only the low-level image features are provided in Table 1, while the confusion matrix for classification results with features from the CNN are provided in Table 2, and the confusion matrix of classification results with the combined features are given in Table 3.

The experiment is repeated for 10 folds. At each run, the F-score values for each class is estimated and their mean is evaluated. Average of the mean F-score values for 10 folds are estimated as shown in Table IV.

Classification results for sample images with all three features are provided in Figure 3. The first image of this Figure is a helicopter image, which is classified as ketch using low-level features, as airplane using features from CNN and it is correctly classified using the combined features. The second image is an airplane image which is correctly classified by low-level features and combined features while misclassified using features from CNN. The last image is a ketch image which is correctly classified using features from CNN and the combined features, while it is misclassified by the low-level features.

Figure 3. Sample images and their classification results with each feature set.

III. CONCLUSION

In this study, low-level image features are combined with features from a CNN and it is observed that the classification performance is improved with the combined features.

If the dataset is not large enough to train a complex CNN model or the system requirements for running a CNN model is not satisfied, then a simple CNN can be employed to extract image features and these features can be combined with low-level features. It is experimentally observed that the classification performance is increased with the combined features.

If system resources are sufficient, the architecture of the CNN can be more complex. But there is always a trade of between the system complexity and the size of the training data required. In this study, visual representation with low-level image features is enhanced with the good representation capability of CNN without avoiding complex architecture.

Apart from monitoring feature extraction from other CNN architectures, visual features such as other MPEG-7 features or Scale Invariant Features (SIFT) [8] can be employed for combining visual features. Alternative

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

methods, instead of concatenation, for combination can be employed in future work.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L.J.V. Gool, “Surf: Speeded up robust features”, in ECCV, 2006, pp. 404-417.

[2] Y. LeCun, K. Kavukcuoglu, and C. Farabet. “Convolutional networks and applications in vision”. In Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, 2010, pp. 253–256.

[3] A. Krizhevsky, I. Sutskever, and G.E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012, pp. 1097-1105.

[4] S. Pan and Q. Yang. “A survey on transfer learning”. Knowledge and Data Engineering, IEEE Transactions on, Volume 22, Issue 10, 2010, pp.1345–1359.

[5] C. Iakovidou, N. Anagnostopoulos, A. Kapoutsis, Y. Boutalis, M. Lux and S. A. Chatzichristofis, “Localizing Global Descriptors for Content Based Image Retrieval. Simplifying the Simple Family of Descriptors”, EURASIP Journal on Advances in Signal Processing», Springer, Volume 2015, Issue 80, 7 September 2015, pp 1-20.

[6] S.A. Chatzichristofis, and Y.S. Boutalis. "CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval.", International Conference on Computer Vision Systems. Springer, Berlin, Heidelberg, 2008, pp. 312-322.

[7] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004.

[8] D. Lowe, “Object Recognition from Local Scale-Invariant Features”, IEEE. ICCV 1999, Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, p 1150.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.