ATTRIBUTE BASED CLASSIFIERS FOR IMAGE UNDERSTANDING

GÖRÜNTÜ ANLAMLANDIRMAK İÇİN NİTELİK TABANLI SINIFLANDIRICILAR

BERKAN DEMİREL

ASST. PROF. DR. NAZLI İKİZLER CİNBİŞ

Supervisor

ASST. PROF. DR. RAMAZAN GÖKBERK CİNBİŞ

Co-Supervisor

Submitted to Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

December 2016

This work named **"Attribute Based Classifiers for Image Understanding"** by **BERKAN DEMİREL** has been approved as a thesis for the Degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING** by the below mentioned Examining Committee Members.

Assoc. Prof. Dr. Ismail Sengor ALTINGOVDE
Head ..... 

Asst. Prof. Dr. Nazlı İKİZLER CİNBİŞ
Supervisor ..... 

Asst. Prof. Dr. Mehmet Erkut ERDEM
Member ..... 

This thesis has been approved as a thesis for the Degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING** by Board of Directors of the Institute for Graduate Studies in Science and Engineering.

Prof. Dr. Salih Bülent ALTEN
Director of the Institute of
Graduate School of Science and Engineering

”Hiç dinlenmemek üzere yola çıkanlar, asla yorulmazlar.”

Mustafa Kemal ATATÜRK

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate Studies in Science of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules.
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

02/12/2016

BERKAN DEMİREL

ABSTRACT

ATTRIBUTE BASED CLASSIFIERS FOR IMAGE UNDERSTANDING

Berkan DEMİREL

Master of Science, Computer Engineering Department

Supervisor: Asst. Prof. Dr. Nazlı IKİZLER CİNBİŞ

Co-Supervisor: Asst. Prof. Dr. Ramazan GÖKBERK CİNBİŞ

December 2016, 80 pages

Attributes are mid-level semantic concepts which describe visual appearance, functional affordance or other human-understandable aspects of objects and scenes. In the recent years, several works have investigated the use of attributes to solve various computer vision problems. Examples include attribute based image retrieval, zero-shot learning of unseen object categories, part localization and face recognition.

This thesis proposes two novel attribute based approaches towards solving (i) top-down visual saliency estimation problem, and, (ii) unsupervised zero-shot object classification problem. For top-down saliency estimation, we propose a simple yet efficient approach based on Conditional Random Fields (CRFs), in which we use attribute classifier outputs as visual features. For zero-shot learning, we also propose a novel approach to solve unsupervised zero-shot object classification problem via attribute-class relationships. However, unlike other attribute-based approaches, we require attribute definitions only at training time, and require only the names of novel classes of interest at test time. Our detailed experimental results show that our methods perform on par with or better than the state-of-the-art.

Keywords: attributes, zero-shot learning, saliency estimation, conditional random fields, semantic word vectors

GENİŞLETİLMİŞ ÖZET

GÖRÜNTÜ ANLAMLANDIRMAK İÇİN NİTELİK TABANLI SINIFLANDIRICILAR

Berkan DEMİREL

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Yrd. Doç. Dr. Nazlı İKİZLER CİNBİŞ

Yardımcı Danışman: Yrd. Doç. Dr. Ramazan GÖKBERK CİNBİŞ

Aralık 2016, 80 sayfa

Nitelikler nesne ve sahnelerin görsel, işlevsel ya da insanlar tarafından algılanabilecek diğer yönlerini tanımlayan orta-düzey semantik bilgileri temsil etmektedir. Son yıllarda, araştırmacıların nitelik kavramına ilgisi giderek artmakta ve bununla birlikte nitelik bilgisi çeşitli bilgisayarlı görü problemlerinin çözümünde sıklıkla kullanılmaktadır. Bu ilginin nedenleri çok çeşitli olmakla birlikte temelde iki neden sayılabilir:

1. Bilgisayarlı Görü alanında çalışılan problemlerde yerel ve genel bağlam bilgisini iyi kodlayabilmesi.
2. Giderek büyüyen veri kümeleri üzerinde işaretleme ve etiketleme yapmanın ortaya çıkardığı zorluklar.

Nitelik bilgisinin problemlerin çözümünde kullanım formları çok farklı olabilmektedir. Örneğin, nitelikler ya da nitelik tabanlı sınıflandırıcı çıktıları, nesneleri ve sahneleri tanımlayan anlamsal öznitelikler olarak kullanılabilmektedir. Bunun dışında nitelik bilgileri, nesneler ya

da sahneler arasındaki ilişkiyi tanımlamak için de kullanılmaktadır. Bu ilişkiler sıfır-bilgi öğrenme *(zero-shot learning)* ya da sınırlı-bilgi öğrenme *(few-shot learning)* gibi problemlerin çözümüne büyük katkı sağlamaktadır.

Bu tez, yukarıda bahsi geçen farklı nitelik kullanma yöntemlerini, iki önemli bilgisayarlı görü probleminin çözümünde kullanmaktadır. Bu problemler:

- Yukarıdan-aşağıya dikkat çeken görsel bölge tespiti
- Denetimsiz sıfır-bilgi nesne sınıflandırma

olarak sıralanabilir.

Dikkat çeken görsel bölge tespiti, görsel verilerde insan gözünün öncelikli olarak odaklandığı bölgeleri bulmayı amaçlayan ve son yıllarda dikkat çeken bir bilgisayarlı görü problemidir. Bilgisayarlı görü problemleri çoğunlukla ön plan öğeleri ya da ön plan öğelerinin birbirleriyle olan etkileşimi ile ilgilenmektedir. Dolayısıyla çoğu zaman bir görüntüde ya da sahnede bulunan arka plan öğeleri yok sayılmaktadır. Bu noktada dikkat çeken görsel bölge tespiti yöntemleri, arka plan öğelerini görüntülerden temizleyerek diğer problemlerin çözümünde ön işlem adımı olarak kullanılabilir.

Dikkat çeken görsel bölge tespiti problemini çözmeye yönelik olarak ortaya konulan modeller alttan-üste ve yukarıdan-aşağıya olmak üzere iki ana grupta toplanmaktadır. Alttan-üste yaklaşımlarda resimlerde mevcut olan yerel ipuçlarından faydalanılarak ön plan öğeleri tespit edilmeye çalışılmaktadır. Bu ipuçları genellikle yoğunluk, renk, doku ya da parlaklık bilgileriyle ilintili olmakta ve modeller bu bilgilerin yerel komşuluklardaki ilişkilerini incelemektedir. Yukardan-aşağıya yöntemler ise spesifik hedef tespiti yapmaya çalışmaktadır ve bu açıdan nesne tanıma problemi ile yakından ilişkilidir.

Bu tez kapsamında dikkat çeken görsel bölge tahmini problemi yukarıdan-aşağıya bir yaklaşımla ele alınmaktadır. Yaklaşım, problemi koşullu rastgele alanlar çizgeleri, nitelik tabanlı sınıflandırıcılar ve seyrek kodlama kullanarak çözmeye çalışmaktadır. Bu amaçla, her bir ön plan nesnesi için ayrı koşullu alanlar çizgesi tanımlanmış ve bu çizgelerde öznitelik olarak nitelik tabanlı

sınıflandıcı çıktıları kullanılmıştır. Ön plan nesnelerini diğer nesnelerden daha ayırt edici ve etkili temsil etmek adına seyrek kodlama teknikleri de çözüme dahil edilmiştir.

Geliştirilen yöntem, öznitelik olarak nitelik tabanlı sınıflandıcı çıktılarını kullanması nedeniyle özgün bir yaklaşımdır. Geliştirilen yöntem Graz-02 veri kümesinde test edilmiş ve alt düzey öznitelik kullanan diğer yöntemlere göre daha başarılı sonuçlar ürettiği gözlemlenmiştir.

Bu tez kapsamında ayrıca, denetimsiz sıfır-bilgi nesne tespiti probleminin çözümüne yönelik sınıf-nitelik bilgisini ele alan yeni bir yöntem önerilmiştir. Son yıllarda bilgisayarlı görü alanında kullanılan veri kümelerinin boyutları muazzam seviyelere ulaşmıştır. Bu veri kümeleri üzerinde nesne işaretlemesi yapmak büyük bir emek ve uzun uğraşlar gerektirmektedir. Bu sebeple veri kümelerinin boyutu arttıkça sıfır-bilgi tabanlı yaklaşımların önemi de artmaktadır.

Sıfır-bilgi yaklaşımlarda amaç, eğitim kümesinde bulunan sınıflar üzerinden elde edilen semantik bilgiyi daha önce hiçbir örneği ile karşılaşılmamış test sınıflarına aktarmak ve bu sınıflara ait resimleri doğru şekilde sınıflandırabilmektir. Sıfır-bilgi yaklaşımlarda semantik bilgi yöntemden yönteme farklılık göstermekle birlikte son yıllarda nitelik bilgisi kullanan yaklaşımlar ön plana çıkmaktadır.

Bu tez kapsamında, nitelik bilgisi sıfır-bilgi nesne ve hareket tanıma problemlerinin çözümüne yönelik aktif şekilde kullanılmaktadır. Ancak diğer yöntemlerin çoğundan farklı olarak nitelik kavramlarının yalnızca görsel değil; yazılı metinlerden elde edilen semantik bilgileri de modele dahil edilmiştir. Geliştirilen yaklaşımda kullanılan hipoteze göre bir nesne sınıfına ait semantik kelime vektörü ile bu sınıfa ait niteliklerin ortalama semantik kelime vektörü arasındaki benzerlik diğer sınıflara ait niteliklerin ortalama kelime vektörlerinden daha fazla olmalıdır. Geliştirilen yöntemde, bahsedilen hipotezi gerçekleştirebilmek ve eğitim kümesinden bu hipoteze yönelik ortaya çıkan semantik bilgiyi test sınıflarına aktarabilmek için doğrusal ve doğrusal olmayan dönüşüm matrisleri öğrenilmektedir.

Yukarıda bahsedilen hipoteze dayanarak geliştirilen yöntem denetimsiz olarak çalışmaktadır. Dolayısıyla, test sınıflarına ilişkin nitelik bilgileri bilinmemektedir. Bu bilgiyi elde etmek

amacıyla eğitim kümesi üzerinde nitelik tabanlı sınıflandırıcılar ayrıca eğitilmektedir. Test kümesinde çalıştırılan nitelik sınıflandırıcılardan elde edilen skorlar, o niteliğin ilgili resim için ne kadar geçerli olduğu bilgisini vermektedir.

Geliştirilen yöntem, geleneksel sıfır-bilgi yöntemlerinin çoğunluğundan farklı olarak test sınıflarına ait herhangi bir bilgiye ihtiyaç duymamaktadır. Yöntem farklı veri kümelerinde değerlendirilmiş ve denetimsiz çalışan en iyi yöntemlerden daha iyi sonuçlar ürettiği gözlemlenmiştir. Ayrıca, denetimli (test sınıflarına ilişkin bilgilere ihtiyaç duyan) yöntemlerle yapılan karşılaştırma sonuçlarına göre bu yöntemlerden daha iyi ya da yaklaşık aynı sonuçlar üretmektedir.

Anahtar Kelimeler: nitelikler, sıfır-bilgi öğrenme, dikkat çeken görsel bölge tespiti, koşullu rastgele alanlar, anlamsal kelime vektörleri

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank to my supervisors Asst. Prof. Dr. Nazli Ikizler Cinbis and Asst. Prof. Dr. Ramazan Gokberk Cinbis for their time, patience and for their valuable guidance at every stage of my research. Without their support, It would have been impossible to write this thesis.

Besides, I would like to thank to my thesis committee members Assoc. Prof. Dr. Ismail Sengor Altingovde and Asst. Prof. Dr. Mehmet Erkut Erdem for reviewing this thesis and giving insightful comments.

Moreover, I am grateful to my all friends and colleagues for their help and good wishes. I would like to thank everyone who wants to make the world a better place without expecting any benefit and who dreams beyond the stars.

Finally, I express my most sincere gratitude to my family. They have supported my decisions and taught me what unconditional love means.

CONTENTS

Page

FIGURES

Page

1. INTRODUCTION

In recent years, researchers have been taking advantage of attributes with a growing interest to solve problems of computer vision. Attributes are mid-level semantic concepts which describe visual appearance, functional affordance or other human-understandable aspects of objects and scenes. They provide practical information about the nature of the classes and are successfully employed in several applications [1–6].

The primary reason for the interest shown to attributes is the power of attributes in defining visual entities, e.g. objects, scenes or their parts, [2, 5, 7–20]. A second reason is the transferability of attributes across object and scene categories [21, 22]. For instance, the attribute *tail* is valid for many animals such as *dog, cat, horse* and *donkey*. This allows accurate representation of novel objects and scenes in terms of attributes with reduced effort.

The aforementioned use of attributes may imply a reduction of effort in terms of number of classifiers to be trained, training data collection or feature extraction complexity. In this thesis, we aim to leverage these advantages of attributes in two problems. First, we propose a top-down visual saliency estimation approach using attribute based visual descriptors. Second, we propose an unsupervised zero-shot learning approach based on learning relationships across attribute and class names. These problems are summarized in further detail in the following.

Top-Down Visual Saliency Estimation

Visual saliency estimation is one of the most actively studied research problems in both image processing and computer vision communities. In visual saliency estimation, the goal is to predict where humans primarily look in images through computational models. Most of these models are based on the widely accepted hypothesis that people tend to focus on regions that contain *foreground* objects [23].

The human brain has reached an eclectic system in the evolutionary process and therefore it has become able to analyze complex scenes in a much shorter time than a second [24]. It is not wrong to say that the process of purifying unnecessary information also takes place in the human brain while trying to handle visual problems. As an analogy, in computer vision problems, irrelevant regions can also be filtered with the aid of visual saliency maps

obtained from images or visual scenes. Therefore, visual saliency estimation approaches can be used as a pre-processing step for many problems such as object detection[25], object recognition[26, 27] or video summarization[28]. Moreover, saliency estimation will become more important with the constantly increasing visual data because it provides a mechanism to examine visual data more quickly[25].

Saliency estimation models can be grouped into two main different categories: (i) bottom-up models (*e.g.* [29–36]) and, (ii) top-down models (*e.g.* [37–44]). Bottom-up saliency estimation can be summarized as identifying visually *interesting* regions using local cues extracted from images. These cues can aim to characterize the color, brightness, intensity or texture of the local images regions, and the relationships between these regions in the local neighborhood context. In contrast, top-down saliency estimation is task-oriented and involves generating pixel-wise maps for instances of target object categories. Therefore, it is closely related to the problem of semantic segmentation.

In our approach, we propose a top-down visual saliency estimation method. In particular, instead of using local image features, we apply attribute classifiers to local image patches. These classifier results are then used as feature vectors to represent objects and their spatial relations in images. In the next stage, a visual dictionary of target classes is initialized with the collected attribute classifier results from training images and K-means algorithm. Then, A Conditional Random Field (CRF)[45] model is learned while the discriminative dictionary is being updated. As a joint learning, more discriminative and compact visual dictionary is obtained. The main advantage of using such a joint learning method is that more accurate classifiers and discriminative visual representations are emerged.

Zero-Shot Learning

Emergence of the gigantic visual data collections that contain billions of images have increased the need for automatic labeling in computer vision tasks. Zero-shot learning aims at solving this problem by classifying images of classes that are not seen before by using limited prior knowledge.

The performance of zero-shot learning methods depends heavily on precise definition of shared information between seen and unseen classes. At this point, many existing works describe and transfer shared information over *attributes* [7, 22].

Most of the methods that use attributes as a source of information for zero-shot learning [22, 46, 47] require the presence of attribute annotations of classes during test time. This means that in such approaches, there is still a need for gathering semantic attribute annotations for unseen classes, and collecting such fine-grained information is a time consuming and error-prone task, which inhibits the scalability of zero-shot learning. Some recent approaches [48–50] leverage additional textual data to alleviate this need. These approaches utilize large amounts of textual resources, together with the labeled training examples for the known classes, to learn a mapping that allows building classifiers based on the names or textual descriptions of unseen classes.

While textual resources, like Wikipedia, can provide rich information for building *semantic* word representations, it is difficult to model *visual* characteristics of unseen classes solely based on such textual data. To address this problem, we propose to use attributes as an intermediate layer between image features and class names, based on the observation that attributes provide a way to link the visual and semantic universe. Importantly, unlike the large body of work on attribute-based zero-shot learning, our goal is to use attributes in an unsupervised way, *i.e.* without requiring manually defined attribute-class annotations for the unseen classes.

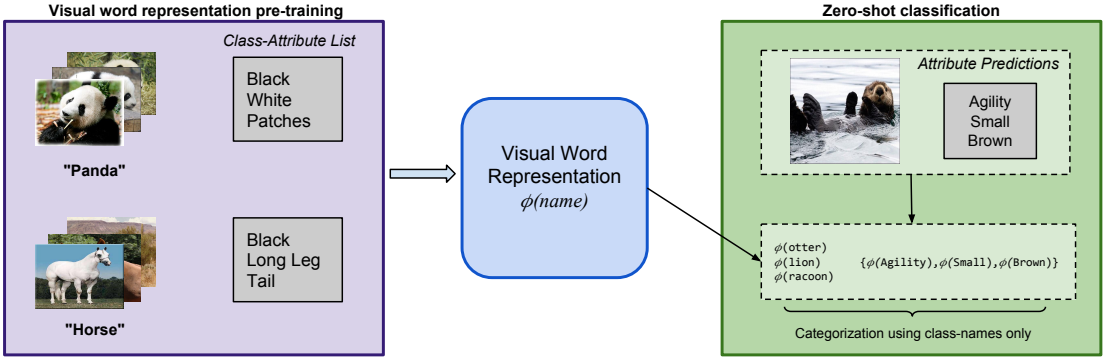


FIGURE 1.1.: Using a set of training classes with known attribute relations, we learn a visually meaningful word representation. The resulting representation is used for zero-shot learning of classes with unknown attribute relations, where attributes form an intermediate layer between visual features and class names.

For this purpose, in this thesis, we learn a visually consistent word representation such that the similarities between class and attribute name representations depict the visual similarity and lead to accurate classifications. More specifically, during training, we learn a transformation matrix from the list of associated attributes and the training class names such that in

the projected space, the word representations of attributes and class names become visually consistent. Noticeably, this word representation learning step is formulated completely over class-attribute relationships of seen classes, without requiring visual examples.

At test time, the learned word representation allows assigning novel images to unseen classes, without knowing the corresponding class to attribute associations or their textual descriptions. The proposed method is summarized on Figure 1.1.

As well as applying our proposed method to standard zero-shot learning datasets, we also apply our model on zero-shot action recognition datasets. In the action recognition datasets, although the videos belong to the same class, their resolutions, frame rates and durations differ from each other. Therefore, the action recognition domain is important to show how the method we developed behave under difficult conditions.

1.1. Major Contributions of This Thesis

This thesis contribute towards a solution for two different computer vision problems. Our main contributions can be summarized as follows:

- A novel approach to improve top-down saliency estimation problem. According to proposed approach, model uses middle-level features to encode visual information in images and their local parts. In this context, model uses results of attribute based classifiers as feature vectors.
- A novel method for learning a visually consistent word representation to improve zero-shot learning problem. The proposed method uses class-attribute relationships to learn semantic embedding matrix which transfers semantic information from seen training classes to the unseen test classes.

In the top-down saliency estimation problem, we use attribute based classifier results as middle-level feature representations of image patches. This approach is based on the [37]. Main difference is that they use only low-level SIFT [51–53] descriptors, whereas we use high-level attribute based descriptors. The number of the attribute based classifiers which are used in this study are less than dimension of SIFT descriptors, so dimensionality of our representation is smaller than the approach which is described in [37].

In the zero-shot learning problem, we propose a novel approach based on textual and visual modalities. In this context, we learn a visually consistent word representation for relating class and attribute combinations purely based on their names. Moreover, as opposed to traditional approaches that leverage attribute information for zero-shot learning, our method does not require attribute annotations or textual descriptions of unseen classes at test time.

1.2. Organization of the Thesis

The rest of the thesis is organized as follows. In the Chapter 2., we give some background information about attributes. Then, we review some related work on top-down and bottom-up saliency estimation techniques, visual attributes, zero-shot learning, textual information and label embedding. In the same chapter, we also talk about datasets and deep models which are used in this thesis.

We examine details of the aforementioned top-down saliency estimation study in Chapter 3. We give the experimental results, compare proposed approach with the previous works and discuss obtained results in the same chapter. In Chapter 4., we give the details of our novel unsupervised zero-shot learning approach. Furthermore, we also give detailed experimental analysis and discuss their impact on literature in comparison with other state-of-the-art approaches.

Finally, we conclude the thesis with a brief summary and a discussion on future research directions in Chapter 5.

2. BACKGROUND AND RELATED WORK

In this chapter, we first briefly introduce attributes and their usage in the computer vision domain. Then, we discuss the most recent works about visual saliency estimation, visual attributes, zero-shot learning, label embedding and textual information. Finally, we introduce related datasets and deep network topologies which are used in this thesis.

2.1. Attributes

Attributes are mid-level semantic concepts which describe visual appearance, functional affordance or other human-understandable aspects of objects and scenes. Visual attributes have been a focus of attention in computer vision community in the last decade, following the pioneering works [7, 21, 54]. Ferrari and Zisserman [54] propose a probabilistic generative model of attributes, where attributes are presumed as patterns of image segments. Lampert *et al.* [21] and Farhadi *et al.* [7] propose methods to annotate objects based on predefined class and attribute relationships. Following these works, attributes have since been successfully used in many applications, *e.g* image classification [55–57], scene recognition [10], face recognition [58], image retrieval [59], action recognition [60, 61], part localization [62], zero-shot learning [21, 22, 46, 49, 63–66].

The use of attributes in computer vision related problems can be basically divided into two main categories: direct and indirect attribute usage. In direct attribute usage, attributes or classifiers which are trained on datasets which contain attribute annotations are used as features of objects and images. In the indirect attribute usage, attributes are used to define high-level semantic relationships between objects or scenes, so these relationships can provide beneficial information to solve some problems like zero-shot learning.

Three different types of attributes have been defined in the computer vision domain. We can summarize these types as follows:

- **Binary Attributes:** This attribute type precisely indicates whether a property belongs to a class or not. They do not contain any probability or comparison between attributes [7, 21].

- **Relative Attributes:** This attribute type works based on relative information between images. They are efficient than binary attributes for enabling semantic relationship between attributes [11]. However, relative attributes are not good at determining attributes in single images.

- **Spoken Attributes:** Spoken attributes are a mixture of binary and relative attributes. They include positive aspects of both methods and use them together. Binary attributes are good at determining attributes in a single image while relative attributes can address comparison of images more successfully [67].

2.2. Deep Convolutional Neural Network (CNN) Topologies

For more than twenty years, number of convolutional neural network architectures have been developed. However, these models had not lead to state-of-the-art results for most problems due to the low calculation capacity of computers. Nowadays, these deep architectures have gained great popularity because of the technological advances occurring in hardware industry, together with advances in deep learning architectures.

AlexNet [68] is a deep neural network architecture which is developed for ImageNet ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) 2012 [69]. It significantly outperformed the other traditional models and popularized convolutional neural networks in the computer vision community. It is developed based on LeNet-5 [70, 71] architecture and it separated from the LeNet-5 model with the following directions:

- AlexNet uses data augmentation techniques that consisted of patch extractions, image translations and horizontal reflections.

- AlexNet prevents the problem of overfitting with implementing dropout layers.

- AlexNet uses the Rectified Linear Unit (ReLU) as non-linearity function.

- AlexNet architecture is trained with using batch stochastic gradient descent.

In the AlexNet architecture 5 convolutional, max-pooling, dropout layers are followed by 3 fully-connected layers. The AlexNet architecture is illustrated on Figure 2.1.

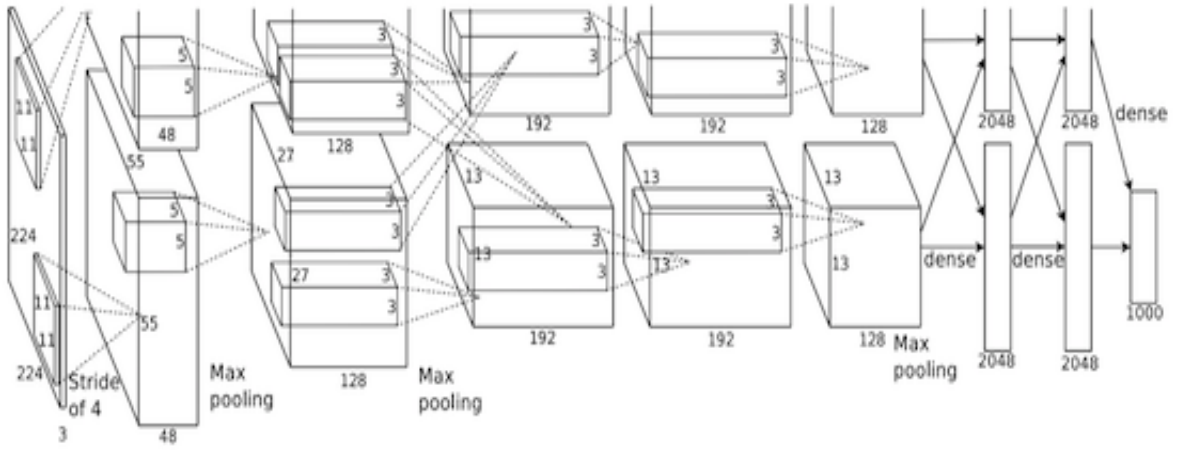


FIGURE 2.1.: An illustration of the AlexNet architecture. (Image is taken from [68].)

After this remarkable development, convolutional neural networks continue to constantly improving the degree of championship in ILSVRC challenges. These developments and innovations are as follows:

- **ZFNet.** It is winner model of the ILSVRC 2013 [72]. It has same architecture with AlexNet, but it varies with some aspects such as receptive fields and filter sizes.

- **GoogleNet.** It is winner model of the ILSVRC 2014 [73]. GoogleNet reduces the computational burden of deep neural networks while obtaining state-of-art performance on ImageNet dataset. Szegedy *et al.* [73] have developed an inception module that reduces the number of parameters between layers, so they had a chance to develop a deeper architecture. Inception Module use 1x1 convolutional filters [74] to reduce the number of features before transferring them to the parallel blocks. The developed inception module is illustrated in Figure 2.2..

- **ResNet.** It is winner model of the ILSVRC 2015 [75]. It has a simple idea: add a learned ”residual” to the its input. In other words, network feeds the output of two convolutional layer and their input to the next layers. In this way, network can learn weight and depth at the same time. A sectional image of this approach is given in Figure2.3..

- **VggNet.** Another architecture that deserves to be mentioned is VggNet [76]. It was the runner-up approach in the ILSVRC 2014. This architecture is the first to use much smaller filters in each convolutional layers. According to the their idea, using small

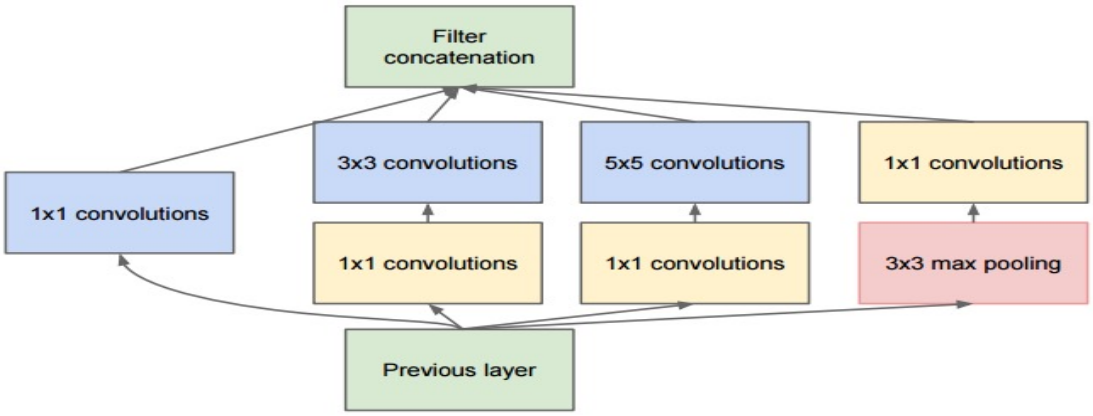


FIGURE 2.2.: An illustration of the Inception Module. (Image is taken from [73].)

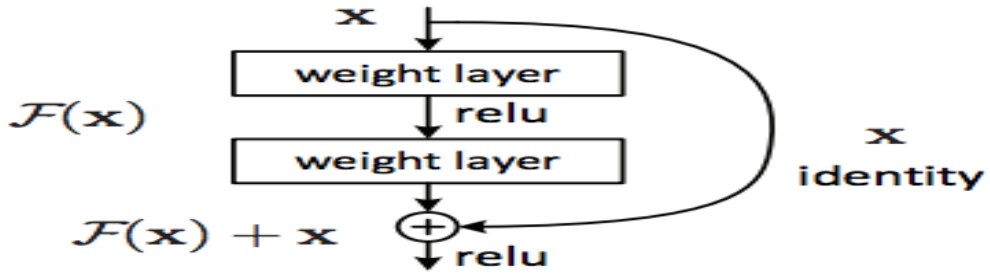


FIGURE 2.3.: Residual learning: a building block. (Image is taken from [75]).

convolutions in sequence can emulate the larger receptive fields. This realistic idea has also been guided for further architectures.

2.3. Visual Saliency Estimation

The goal of saliency estimation is to identify the conspicuous objects in a scene or an image. Recently, this research field has gained significant interest in computer vision community because it provides a mechanism to filter out unwanted information in a visual scene. Recent approaches can be divided into two main categories: top-down approaches [37–44] and bottom-up approaches [29–36].

Yang and Yang [37] learn jointly discriminative dictionary and conditional random fields on patch-based representation. This work shows that the obtaining more discriminative dictionaries for foreground objects improve performance of approaches. Kocak *et al.* [38] convert patch-based representation of [37] to the superpixel-based representation. They show that

superpixel based representation can handle boundary information more reliable than patch-based approaches.

Top-down visual saliency models are generally trained in a supervised setting, but Cholakkal *et al.* [77] propose a weakly supervised framework which contains only binary image labels that indicate the presence or absence of a target object in an image. This proposed novel approach is based on R-ScSPM framework which enables selection of representative image patches for contextual saliency. Cholakkal *et al.* [78] also revise their approach with replacing sparse codes of SIFT features with CNN features.

Liu *et al.* [79] formulate salient object detection as an image segmentation problem. For this purpose, they prepare multi-scale contrast and spatial color distribution features and combine them with Conditional Random Fields efficiently.Borji and Itti [80] combine different saliency estimation maps from different color spaces. They show that some objects in an image are more salient in RGB color space, while saliency detection works better in Lab color space for some others. Jiang *et al.* [81] integrate three different and important visual cues(uniqueness, focusness and objectness) in their model. *Uniqueness*, one of the three visual cues, is used to encode visual contrast information, *focusness* is used to encode focus information about foreground objects and *objectness* is used to encode integrity information of salient region.

Zhang *et al.* [82] try to characterize images with binary images of them that are collected from different color channels of image with different random threshold values. They try to find saliency map with using topological structure of the binary images according to the figure-ground segregation principle of Gestalt. Zhu *et al.* [33] try to prepare a novel approach to estimate robust background models instead of foreground object modelling. Then, low-level cues are used with these background models to estimate visual saliency regions.

Kim *et al.* [34] shift low dimensional RGB color space to high-dimensional feature space. Their approach show that foreground objects are more distinguishable in the new high-dimensional feature space. Erdem and Erdem [32] use patch based covariance descriptors to create bottom-up saliency models. Main contribution of this work is solving feature map extraction and feature integration steps in a single-shot.

2.4. Zero-Shot Learning

Zero-shot learning is a process of transferring information from seen classes to unseen classes. Most of the existing work use attributes by means of knowledge transfer between seen and unseen classes. Lampert *et al.* [21, 22] are among the first to use attributes for zero-shot learning. They propose direct attribute prediction (DAP) and indirect attribute prediction (IAP) methods. In both of the approaches, attribute and class relations are provided explicitly. In DAP method, a classifier is learned for each attribute, whereas in IAP, a classifier is learned for each class.

Al-Halah *et al.* [46] apply attribute label propagation on object classes, such that attributes are used at different abstraction levels, making it suitable for fine-grained zero-shot classification problem. Rohrbach *et al.* [47] prepare similar hierarchical method with Al-Halah *et al.* [46], but they use only class taxonomies. Deng *et al.* [83] introduce Hierarchy and Exclusion (HEX) graphs as a standalone layer to be used on top of any-feedforward architecture for classification. Though they did not design these graphs for zero-shot learning, they show that HEX graphs can also be built using exclusion and overlap relations between classes and attributes.

Jayaraman and Grauman [84] propose a random forest approach to handle error tendencies of attributes. In their approach, a decision tree is built for every unseen class using the attribute signatures, and attribute classifier predictions are used as inputs to those classifiers. Romera *et al.* [85] develop two linear layered network to handle relations between classes, attributes and features. The first layer is learned to describe the relationships between features and attributes, and the second layer models the relationships between attributes and classes, and its weights are fixed using the prescribed attribute-class predicate matrices.

Zero-shot learning is used not only in object classification but also in action classification problems [66, 86]. Jain *et al.* [66] use convex combination of action and object similarities to assign action labels of unseen videos. Their approach does not need any specification about attribute-class or class-scene mapping, and the only thing that needs to classify the unseen videos is semantic embedding space.

An important limitation of the aforementioned methods is their dependency on the attribute signatures of the test classes. To scale these approaches to additional unseen classes, the attribute signatures of those new classes need to be provided explicitly. Our method alleviates

this need with learning a word representation that allows associating classes and attribute combinations based on their names.

Label Embedding. Label embedding techniques [87, 88] are being recently introduced for zero-shot learning [49, 65, 89–92]. Akata *et al.* [65] propose attribute label embedding (ALE) method that uses web scale annotation by image embedding (WSABIE) technique [93] as an infrastructure. Different from WSABIE, ALE uses attributes as a side information. Recently, Akata *et al.* [49] improve this ALE method by using embedding vectors that were obtained from huge text corpora instead of using hand-crafted class-attribute embedding.

Another label embedding approach is Frome *et al.*'s method [90] where the authors change winner network of the Krizhevsky *et al.* [68] with zero-shot learning settings. They change the softmax layer with a kind of transformation layer, mapping images into a rich semantic embedding space. Similarly, Norouzi *et al.* [89] also change the network of [68], but in order to handle combination of semantic embeddings, they add one more layer on softmax layer instead of changing it. Wang and Chen [64] prepare a novel approach based on a stagewise bidirectional latent embedding for zero-shot classification. In the bottom-up stage, latent embedding space is created via data of training classes to guide to embed semantics of unseen classes. In the top-down stage, semantic representation of unseen classes are projected to the previously learned semantic space with using Semi-supervised Sammon Mapping [94].

The closest work to ours is the ALE method [49], which uses a pre-trained word representation as the output embedding. Therefore, ALE, from which our work is inspired from, allows zero-shot learning of unseen classes based on their names, while implicitly assuming that word representations provide a visual similarity measure across class names. However, as the word representation models are learned from text corpora, the embeddings they provide are likely to be dominated by the semantic relationships across classes, rather than their low-level visual differences.

To address this limitation, we propose to use attributes as an intermediate layer that connects the visual features and class names, and learn a more visually consistent word representation for constructing this connection. Crucially, unlike the previously proposed attribute-based zero-shot learning approaches, our model requires class-attribute relationships only at training time and not at test time, which we believe is an important step towards large-scale zero-shot classification.

Akata *et al.* [95] propose a novel approach to jointly embed language representations and semantic visual part of images to solve fine-grained zero-shot recognition problem. In this approach, they use textual NAD (Noun-Attribute Difference) to create semantic embedding spaces. Xian *et al.* [96] improve ALE [49] approach with using multiple visual embedding spaces. According to their intuition, different embedding spaces encode different visual characteristics of object classes.

Domain Adaptation with Textual Information. Ba *et al.* [50] propose a method to solve zero-shot fine-grained object classification problem with combining MLP and CNN networks by applying a kind of transformation function on results of networks. Networks handle text based information acquired from Wikipedia articles and visual based information coming from images, respectively. Zhang and Saligrama [97] prepare a novel method to solve zero-shot learning problem with semantic similarity embedding. They learn an embedding technique to represent target classes with histograms of the source classes. Another interesting direction is explored by Elhoseiny *et al.* [48], where they extract information and build classifiers based directly on textual corpus that is accompanied with images. In such approaches, noise within the articles should be carefully handled to extract informative bits.

2.5. Distributional Word Representations

In addition to extracting textual information with standard representations, word embedding techniques [98–100] and structured textual sources [101–104] are becoming very popular [49, 65, 89, 90], due to the powerful vector space representations where the distances can be meaningfully utilized.

word2vec [98, 99] and GloVe [100] methods, among these techniques, have become more popular in recent years. word2vec is a two-layer artificial neural network that works on textual data and task of this neural network is to predict the context of textual data. For this purpose, two different network architectures have been created named CBOW and Skip-Gram. In the CBOW approach, the order of words is not preserved and same projection layer shared for all words, so they are projected into the same position. Because of this, the architecture is named as continuous bag-of-words model. Skip-Gram model is not predict the current word by looking at the contextual information, it increases the possibility of

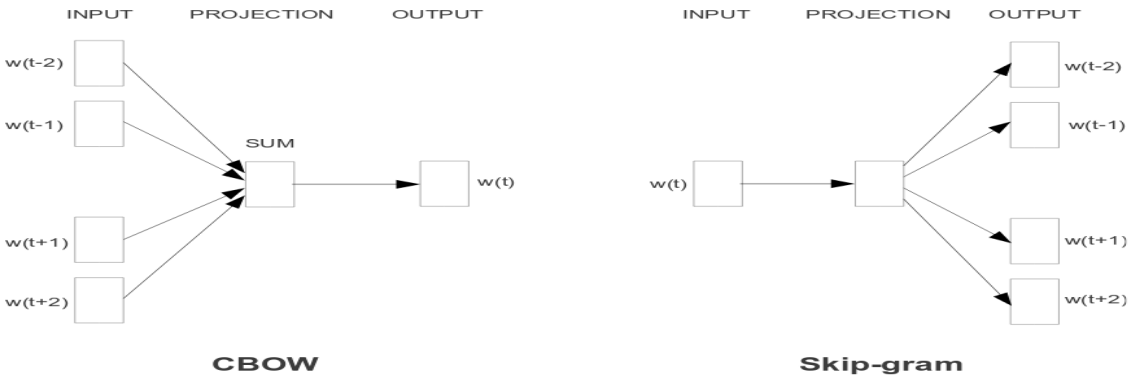


FIGURE 2.4.: Two different word2cec architectures. (Diagrams are taken from [105].)

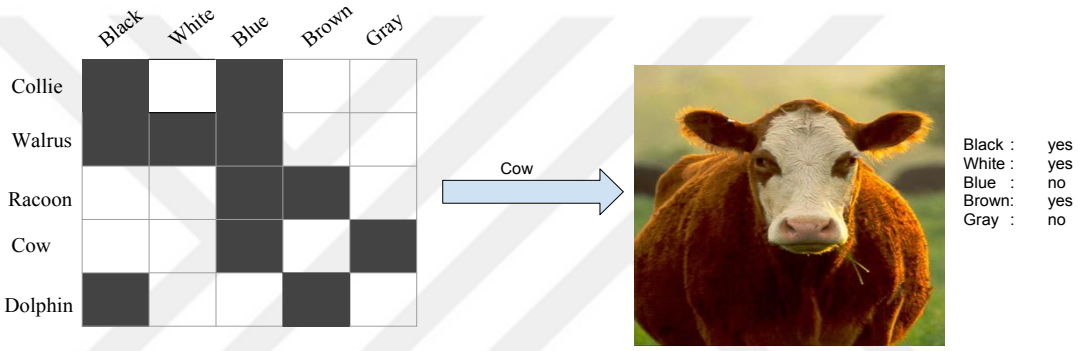


FIGURE 2.5.: Attribute details of AwA dataset.

coexistence scores of a word based on another word in the same sentence. The models described above are shown on Figure 2.4.

GloVe [100] is a count based approach, unlike the word2vec methods. According to the GloVe approach, a co-occurrence counts matrix is generated and dimensionality reduction operation is applied on this matrix. It is attempted to obtain lower-dimensional word vector representations that explain the variance in the high-dimensional word vectors.

2.6. Attribute Datasets

Animals with Attributes Dataset. AwA dataset [22] contains 30,475 images of 50 different animal classes. 85 per-class attribute labels[1] are provided in the dataset. The authors also provide a predefined split for zero-shot learning. In this setting, 40 animal classes are marked

[1] http://attributes.kyb.tuebingen.mpg.de/

FIGURE 2.6.: Visual examples from AwA Dataset

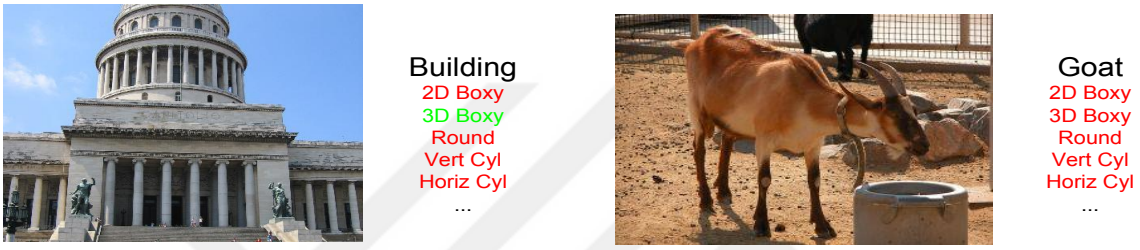


FIGURE 2.7.: Visual examples from aPaY Dataset.

for training and 10 classes for testing, respectively. Some images of this dataset is shown on Figure 2.6.

In this dataset, per-class attribute values are defined as binary labels. Some part of these binary attributes are shown on Figure 2.5.. In this figure, black boxes represent absence of attributes and white boxes represent presence of attributes. The defined class and attribute correlations of AwA dataset is obtained from Osherson's class-attribute matrix [106, 107].

aPaY Dataset. aPaY dataset [7] is formed of images obtained from two different sources[2]. aPascal (aP) part of this dataset is obtained from PASCAL VOC 2008 images [108]. This part contains 12,695 images of 20 different classes. The second part, aYahoo (aY), is collected using Yahoo search engine and contains 2,644 images of 12 object classes which are completely different from aPascal. Images are annotated with 64 binary per-image attribute labels. According to the predefined zero-shot learning settings on this dataset, aPascal part is used for training and aYahoo part is used for testing. Some images of this dataset is shown on Figure 2.7.. In this figure, valid attributes are marked with green color, invalid attributes are marked with red color.

[2]http://vision.cs.uiuc.edu/attributes/

TABLE 2..1: Table of Characteristics of UCF101. Information is taken from [109].

Actions	101
Clips	13320
Groups per Action	25
Mean Clip Length	7.21 sec
Min Clip Length	1.06 sec
Max Clip Length	71.04 sec
Frame Rate	25fps
Audio	Yes (51 actions)

FIGURE 2.8.: Some frame examples from UCF-101 Action Recognition Dataset. (Image is taken from [109].)

2.7. Action Recognition Datasets

UCF101 - Action Recognition Dataset. UCF101 action recognition dataset [109] contains 101 action classes, over 13320 clips and 27 hours of video data. The authors provide 115 binary scene attributes[3] for each of the action classes. The videos are downloaded from YouTube [4], so this dataset consist of user-uploaded videos which might contain cluttered background and non-regular camera motions.

The relevant features of videos in the UCF101 database are shown in Table 2..1. Moreover, some frames for this dataset is shown in Figure2.8.

HMDB51. HMDB [110] is a large scale motion dataset which are mostly collected from movies and a small part of it created from publicly available platforms such as Google and

[3]http://crcv.ucf.edu/THUMOS14/download.html
[4]www.youtube.com

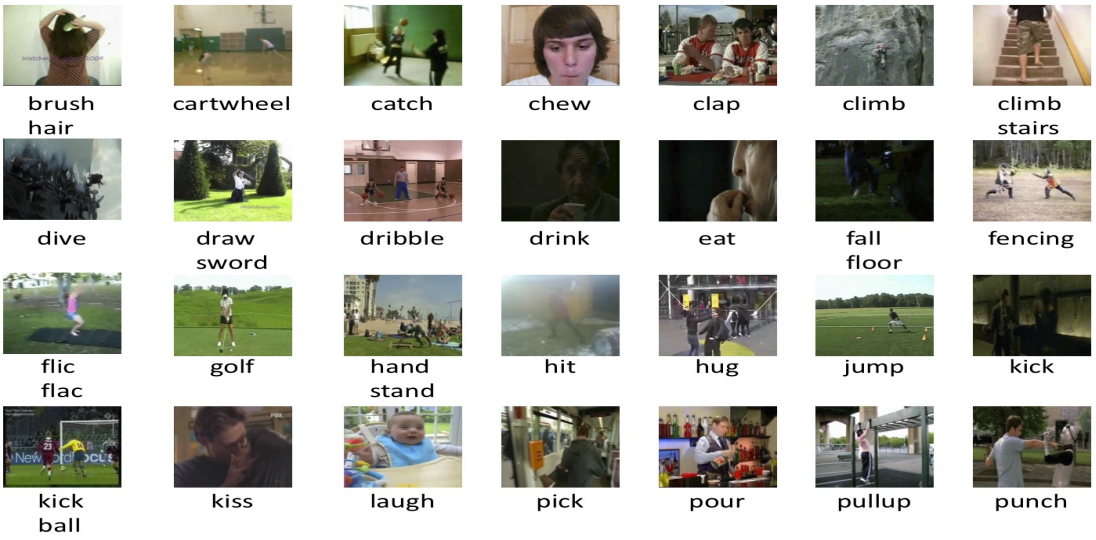


FIGURE 2.9.: Some frame examples from HMDB51 Action Recognition Dataset. (The image is taken from [110].)

TABLE 2..2: Table of Characteristics of UCF-Sports. Information is taken from [111].

Actions	10
Clips	150
Min Clip Length	2.20 sec
Max Clip Length	14.40 sec
Frame Rate	10fps
Resolution	720x480
Max. number of clips per class	22
Min. number of clips per class	6

YouTube. HMDB dataset consists of 51 different action categories and all of these actions are related to human. The dataset contains 6849 clips of action categories and each containing a minimum of 101 clips[5]. Some frames for this dataset is shown in Figure2.9.

UCF Sports Action Recognition Dataset. UCF Sports Action Recognition Dataset [111] is formed of videos from various sport actions which are featured from television channels such as the BBC and ESPN. This dataset contains total of 150 videos of 10 different sport action classes.

The relevant specs of videos in the UCF Sports database are shown in Table 2..2. Furthermore, some frames for this dataset is shown in Figure2.10..

[5]http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

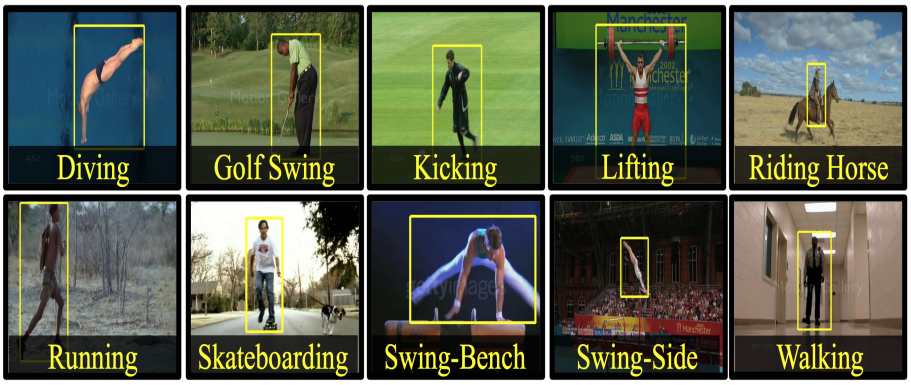


FIGURE 2.10.: Some frame examples from UCF-Sport Action Recognition Dataset. (Image is taken from [111].)

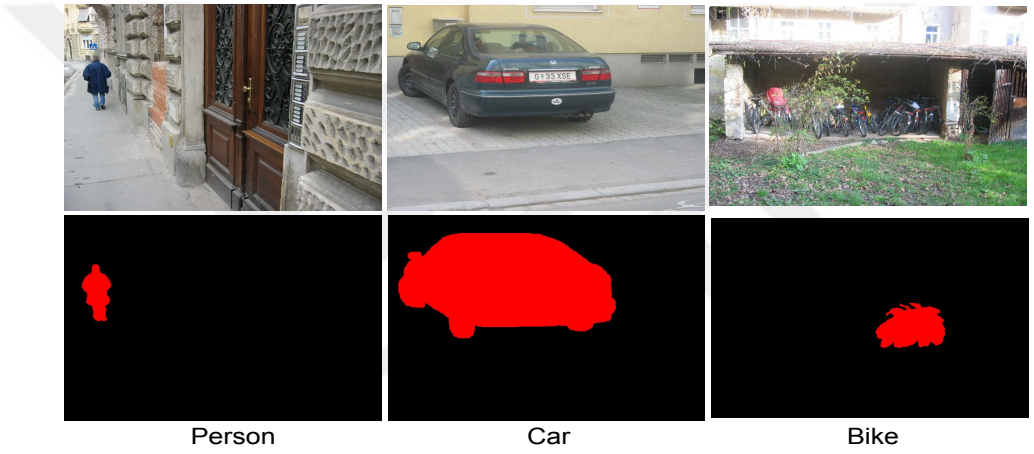


FIGURE 2.11.: Some visual examples from Graz-02 Dataset

2.8. Saliency Estimation Datasets

Graz-02 Dataset. Graz-02 Dataset [112, 113] contains images from three foreground categories (bikes, people, cars) and one counter-class(background) category. It[6] contains 365 images with bikes, 311 images with persons, 420 images with cars and 380 images not containing one of these objects. Moreover, dataset provides ground truth data for 300 images of each category. It is given in terms of pixel segmentation masks with values between 0 and 255 where pixels with 0 denote the object in the image. Some images of this dataset is shown on Figure 2.11.

[6]https://lear.inrialpes.fr/people/marszalek/data/ig02/

3. TOP-DOWN VISUAL SALIENCY ESTIMATION

In this chapter, we introduce our solution on top-down visual saliency estimation via attribute based classifiers and conditional random fields. In this work, we use results of attribute based classifiers as visual feature vectors and then we jointly learn Conditional Random Fields (CRFs) and discriminative dictionaries from these feature vectors.

Rest of this section is as follows. In the Section 3.1., we talk about the proposed approach. Moreover, we give detailed information about attribute based classifiers and their usage, conditional random fields and sparse dictionaries in the same section. In the following section, Section 3.2., we evaluate our method on benchmark dataset and compare it with another related approach. In the Section 3.3., we discuss our results and talk about future of the this approach.

3.1. Approach

We have prepared an approach to improve visual saliency estimation problem in a top-down manner. The developed approach has been built on the state-of-the-art method of Yang and Yang [37] which use SIFT[51–53] descriptors to encode visual features of images. Unlike their work, our model uses ready-to-use attribute based classifiers to predict visual salient regions in images.

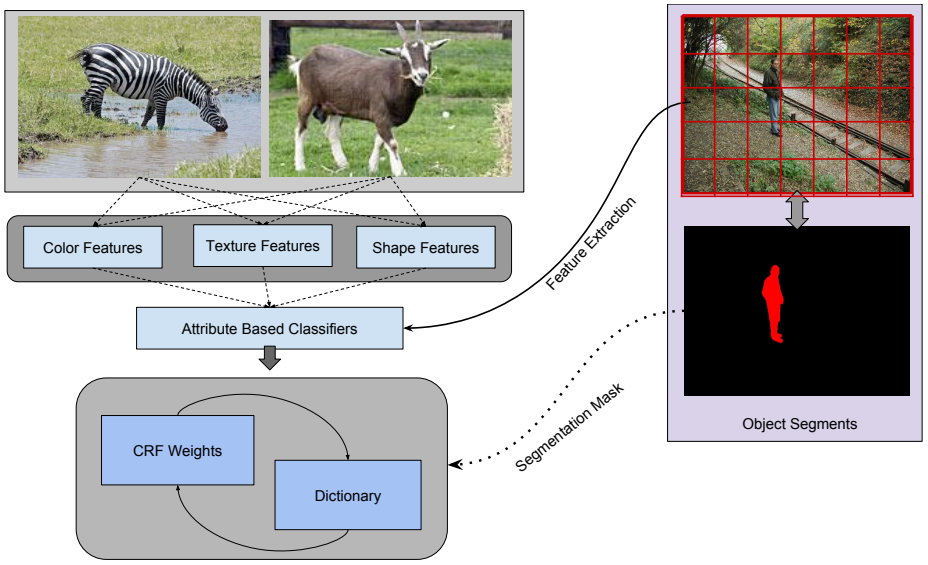


FIGURE 3.1.: Top-Down Visual Saliency Estimation Framework

The method we have developed consists of multiple steps. Initially, we extract equal size visual patches of training images and run attribute based classifiers on these patches. Then, these classifier results are used as feature vectors and these features are fed into *K-Means* clustering algorithm [114] to learn discriminative dictionaries. After that, Conditional Random Fields are created and weights are initialized through the dictionaries and the segmentation masks of the images in the training set. At the same time, more discriminative and compact dictionary is obtained according to the CRF results. These steps are repeated with a certain number of iteration and parameters are tuned. Hence, we jointly learn CRF weights and discriminative dictionaries from training images and ground truth segmentation masks.

Proposed method is also summarized on the Algorithm 1 and also visualized on Figure 3.1.. Moreover, we can summarize the related figure and algorithm of the method as follows:

1. Learn optimized weights of CRFs and compact visual dictionary in the training process.
2. Extract visual patches from test images and run ready-to-use attribute based classifiers on them at the test stage.
3. Try to generate visual saliency maps of images with using learned models which contain attribute features, learned CRFs and visual dictionaries.

CRF and Dictionary Learning

CRFs provide an opportunity to use local context information between neighbour visual patches. Besides, sparse coding also allows more compact and discriminative representation of these patches. However, the approach in [37] is not only a simple combination of CRFs and sparse representations but this approach also uses stochastic gradient descent algorithm to learn CRFs and sparse dictionaries jointly.

Our solution also use the integrated learning aspect of [37]'s approach. In our work, p-dimensional patches which are obtained from images are denoted by $X = [x_1, x_2, x_3, ..., x_m]$ and binary labels which imply the absence of target object on these patches are denoted by $Y = [y_1, y_2, y_3, ..., y_m]$. The developed approach tries to learn a dictionary ($D = [d_1, d_2, d_3, ..., d_k]$)

Algorithm 1 Saliency Estimation Algorithm

$Te \leftarrow$ ListOfTestImages(D)
$Tr \leftarrow$ ListOfTrainingImages(D)
for all $i \in Tr$ **do**
    $P \leftarrow$ PatchesOnImage(i)
    $C_i \leftarrow$ AttributeClassifiers(P)
end for
$D \leftarrow$ GetDictionary(C)
for all $i \leq T$ **do**
    $CRF_{weights} \leftarrow$ UpdateCRFParameters(D, C)
    $D \leftarrow$ UpdateDictionary($CRF_{weights}, C$)
end for
for all $i \in Te$ **do**
    $P \leftarrow$ PatchesOnImage(i)
    $C_i \leftarrow$ AttributeClassifiers(P)
    $SaliencyMap_i \leftarrow$ GenerateSalienyMap($CRF_{weights}, C_i, D$)
end for

which can distinguish target objects from background with the help of the previously mentioned image patches and target object labels.

In order to obtain more informative representations of the target objects in the dictionary, sparse representation of each visual part is obtained via Eq. (1) and ℓ_1-regularization function. Each X patch is represented with $S(X, D)$ hidden variables after sparse coding process expressed in Eq. (1):

$$S(X, D) = \underset{S}{\operatorname{argmin}} \frac{1}{2} \parallel X - DS \parallel^2 + \lambda \parallel S \parallel_1 \tag{1}$$

In this equation, first term tries to provide better representation of X patches and second term determines sparsity level. Moreover, sparse penalty value coefficient is denoted by λ.

The hidden variables which are obtained from X patches with Eq. (1) can be represented as follows:

$$S(X, D) = \Big[S(x_1, D), S(x_2, D), S(x_3, D)...S(x_m, D)\Big] \tag{2}$$

Therefore, visual information are transferred from dictionaries to the hidden variables, so CRF models can be formed as follows:

$$P(Y \mid S(X,D),W) = \frac{1}{Z} e^{-E(S(X,D),Y,W)} \tag{3}$$

Here, partition function is denoted by Z and energy function is denoted by $E(S(X,D),Y,W)$. Moreover, weight vector is denoted by W. Eq. (3) calculates the probability value of Y label according to the given $S(X,D)$ and W values. During the training, $S(X,D)$ and W values are optimized to provide compatibility between ground truth masks and Y labels.

The developed model also works patch based in the test stage. If one part of the image contains target object, probability of neighbour patches about target object containing is increasing, so context information has impact on pairwise potential of energy function. Hence, target information on the patches can be calculated as follows:

$$P(Y_i \mid s_i, w) = \sum_{Y_{N_{(i)}}} P(y_i, Y_{N_{(i)}} \mid s_i, w) \tag{4}$$

Here, $Y_{N_{(i)}}$ represents 4-way neighbourhood of related patch on the graph. Finally, salient region probability of a patch can be calculated as follows:

$$u(s_i, w) = P(y_i = 1 \mid S_i, w)$$

Detailed explanations about internal and pairwise potential of energy minimization function and sparse coding can be found in the [37].

Attribute Based Classifiers

In this work, we use ready-to-use attribute based classifiers which are trained by Farhadi *et al.* [7]. According to the their approach, a binary SVM classifier [115] is learned for each concrete attribute by using different kind of hand-crafted features which are intended to obtain color, texture, and shape information of objects. In this work, color and texture information is used to encode material information; visual words is used to encode part information and edge structure is used to encode shape information.

Texture descriptors are calculated for each pixel on texton filter bank and cluster centers are determined with the aid of the k-means *(k=256)* algorithm. Each pixel is labeled with

the label of the closest cluster center. Histogram of Gradients (HOG) [116] information are obtained via spatial pyramid using 8x8 blocks and 4 pixel step size. Then, cluster centers are determined with the k-means algorithm *(k=256)* and HOG descriptors are quantized into these cluster centers. Edges are determined with Canny Edge Detection algorithm [117] and quantized into 8 cluster centers. Color descriptors are run on each pixel and quantized into 128 cluster centers with k-means algorithm. Moreover, color values are obtained from the LAB color space [118].

In the next stage, aforementioned features which are extracted from images are combined as a single feature vector to be fed into Support Vector Machines. Feature selection is carried out with ℓ_1-regularized logistic regression in order to get accurate attribute based classifiers. In this context, each class which is associated with related attributes is examined and features for related attribute are determined with the logistic regression. After that, pooling operation is applied on features which are obtained from related classes.

3.2. Experiments

In this section, we employ the proposed top-down saliency model on the Graz-02 Dataset [112, 113] and compare results with [37]. Our evaluation criteria is pixel-level precision rates at equal error rates (EER). In the following sub-sections, we give detailed information about experimental setup and discuss the observed results.

Experimental Setup

In our approach we learn a different graph model for each object class separately, so positive labeled images of the training set differ for each model. Moreover, we choose first 300 images of each classes which are defined in the Graz-02 dataset because ground truth object segmentations of other available images are not provided. In this context, we use odd-numbered images for training process, other images for evaluation.

Graz-02 dataset contains images from 4 different categories: *person, car, bike and background*. Images belonging to the background class are used as negative samples in each graph model, so we use 150 odd-numbered images of related object classes as positive training set and 150 odd-numbered images of background class as negative training set.

We extract 999 different patches from each image using a 64x64 pixel sub-window size and a 16-pixel window shift amount. Then, we run attribute based classifiers on them and 64 different binary attribute based classifier results are obtained for each patch. These classifier results are then used as feature vectors to represent objects and their spatial relations in images.

In the next stage, a visual dictionary of target classes is initialized with the collected attribute classifier results from training images and K-means algorithm. In the following stage, a Conditional Random Field (CRF) [45] model is learned while the discriminative dictionary is being updated. As a joint learning problem, more discriminative and compact visual dictionary will be obtained according to the current CRF weights and at the same time CRF weights will also be updated. Iteratively, model determines optimal, compact and discriminative dictionary and proper CRF weights in 20 iteration. CRF weights are learned with 512 visual words. Without of this, we determine λ parameter as 0.15, we set the initial learning rate as 1e-3 and weight penalty value (γ) as 1e-1[1].

Test images are also needed to be represented patch-based, so we extract visual patches from test images and run ready-to-use attribute based classifiers on them. Then, we evaluate the saliency maps of the test images with the learned CRF weights and discriminative dictionaries.

Results

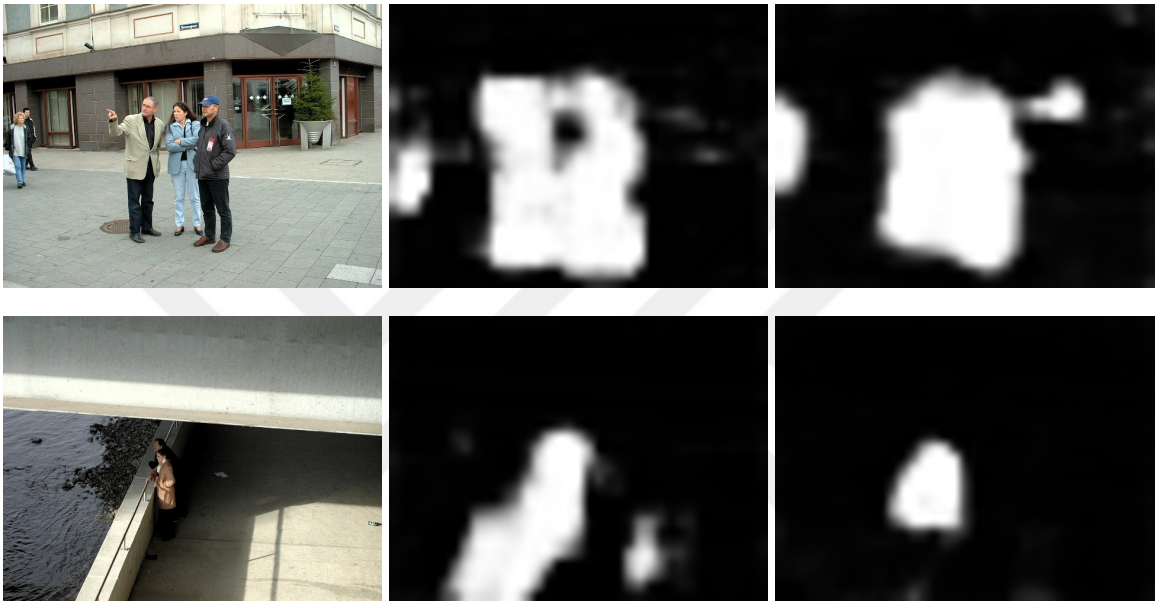
We compare results of the our approach with using pixel-level precision rates at equal error rates (EER). According to the our observations and results, our models find salient regions more accurate than [37] in all object classes. Classification results are given on the Table 3..1. The main reason of this difference is feature representation. [37] use SIFT descriptors to learn discriminative dictionaries and CRF weights. However, we use attribute based classifiers to learn dictionaries and CRF weights.

Some visual results of methods are given on Figure 3.2., 3.3. and 3.4. According to quantitative and qualitative results, we observe that the high-level prior information that are incorporated using attributes lead to better saliency estimation performance. Attribute based classifiers are middle-level features but low-level features are used in the infrastructure of

[1]These empiric parameters are obtained from previously published comparison method [37].

TABLE 3..1: Classification rates at EER.
Parameters: λ = 0.15, γ = 1e-1 and P_0 = 1e-3

	Person	Car	Bike	Average
Yang [37]	50.6	54.3	58.2	54.5
Our Method	**58.7**	**60.6**	**64.9**	**61.4**



(d) Input (e) [37] (f) Our Method

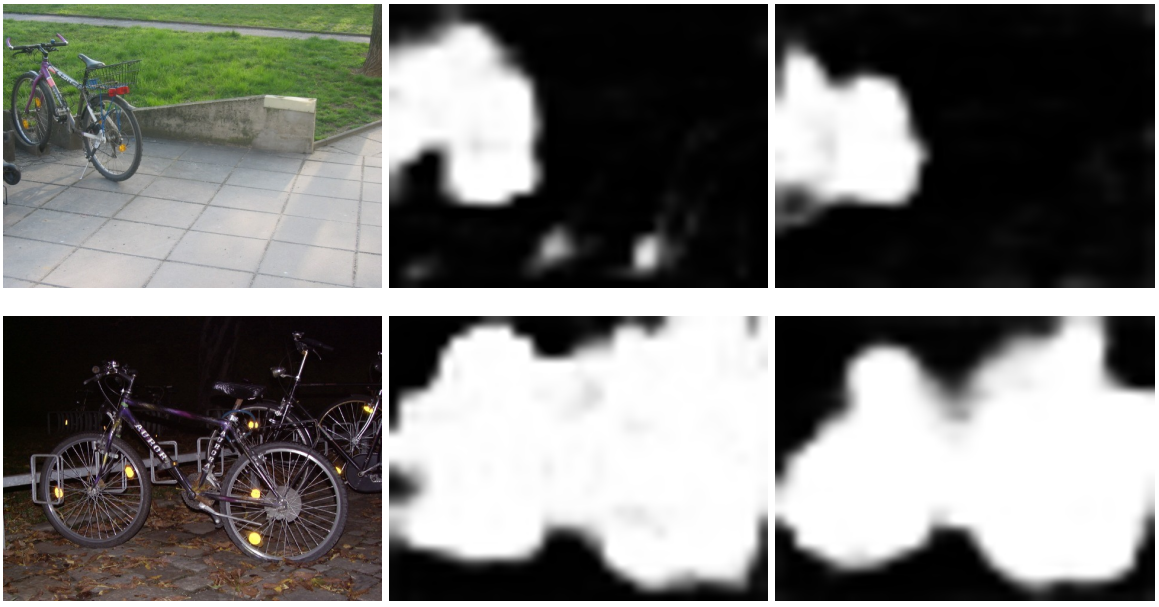
FIGURE 3.2.: Example visual saliency maps obtained for the person class.

them. The main lesson learned from here is correct and compact combination of low-level features might generate more successful results. Besides, Kocak *et al.* [38] achieved an average of **70.17%** success on this dataset. The main reason for this success difference is that the images are represented with different techniques. [38] use a superpixel based representation to handle object boundary information more successfully. Moreover, Cholakkal *et al.* [77] achieved an average of **72.16%** success on Graz-02 dataset because their framework enables selection of more representative patches for contextual saliency. However, our aim here is not to compete with state-of-the-art, but to show that top-down saliency estimation benefits from using attribute classifiers.

However, we do not use target specific attribute classifiers, which are trained on 20 different classes, and this situation affects the success rate of the saliency maps. Figure 3.5. shows that body region of *dog* is perceived as a target object in our method because dog object close

(d) Input (e) [37] (f) Our Method

FIGURE 3.3.: Example visual saliency maps obtained for the car class.



(d) Input (e) [37] (f) Our Method

FIGURE 3.4.: Example visual saliency maps obtained for the bike class.

to the person object and our attributes are not object specific. Hence, the *dog* object is taken into account as a target object in our approach.

FIGURE 3.5.: An example failure case in our attribute-based top-down saliency model.

3.3. Conclusions

In this work, we prepared a method to improve top-down saliency estimation problem. According to our approach, we use results of attribute based classifiers as feature vectors. Experimental results show that attribute based classifiers generate more reliable and compact results than SIFT descriptors.

Moreover, attribute based classifiers produce these accurate results with smaller feature dimensionality. This situation shows that attribute based classifiers are useful and distinctive than SIFT descriptors because they expose high-level object properties by using different kind of attributes and by additional information.

Another noteworthy point is that attribute based classifiers can be used more efficiently. Our classifiers are trained on aPascal dataset and it contains 64 different attributes that are belong to 20 different object classes. Some of these object classes are irrelevant to our target objects, so dataset contains irrelevant attributes for our purpose. In our approach, we use all attribute results regardless of the type of the target object, but we can use attribute based classifiers according to the their relationship with target objects.

4. ZERO-SHOT OBJECT CLASSIFICATION

In this chapter, we introduce our novel approach on unsupervised zero-shot object classification. In this work, we learn a visually consistent word representation such that the similarities between class and attribute name representations depict the visual similarity and lead to accurate classifications. More specifically, during training, we learn a transformation matrix from the list of associated attributes and the training class names such that in the projected space, the word representations of attributes and class names become visually consistent.

Rest of the this section is as follow. In the Section 4.1., we talk about our proposed approach. Moreover, we give detailed information about our hypothesis on semantic relationship between classes and attributes. In the following section, Section 4.2., we evaluate our method on benchmark datasets and compare another related approaches on zero-shot learning. In the Section 4.3., we discuss our results and talk about future of the this approach.

4.1. Approach

In this section, we present the details of our approach. First, we present the main formulation, and explain how the model is used for zero-shot classification on novel test images for unseen classes, without any training images or class-attribute relationships. We also discuss two different application of our approach: (i) zero-shot object recognition, and (ii) zero-shot action recognition.

Learning visually-consistent word representations

Our main goal is to learn a model using which we can achieve zero-shot classification solely based on class names. For this purpose, we define a discriminant function f based on a given class name c and a given set of attributes A:

$$f(c, A) \quad = \quad \sigma\left(\phi(c), \Phi(A)\right) \tag{5}$$

where $\sigma(x, y)$ is the cosine-similarity function, and $\phi(c)$ is the vector representation of the class name c. Similarity, $\Phi(A)$ is the representation of the set of attributes, and is defined as

the average of per-attribute name representations:

$$\Phi(A) = \frac{1}{|A|} \sum_{a \in A} \phi(a). \tag{6}$$

It can be seen in Eq. (5) that the discriminant function is built upon the similarity between the name of a given class and a set of attributes in an image, where attributes provide an intermediate layer in relating the visual space of the image features and the semantic space of class names. Therefore, the classification performance, especially for zero-shot learning, strongly depends on the word representation ϕ. That is, we need to learn a visually meaningful representation ϕ such that we can accurately transfer the visual knowledge from the seen classes to unseen classes.

Our starting point is the principle that the similarity between a class c and the set of associated attributes A_c should be higher than its similarity to any other attribute combination A', *i.e.*:

$$f(c, A_c) \geq f(c, A'), \qquad \forall A' \neq A_c \tag{7}$$

We can extend the constraint in Eq. (7) such that the margin is forced to be proportional to the difference between a pair of attribute combinations:

$$f(c, A_c) \geq f(c, A') + \Delta(A_c, A'), \qquad \forall A' \neq A_c \tag{8}$$

where $\Delta(A, A')$ measures the discrepancy between the two attribute combinations. In our experiments, we use the average Hamming distance between the pair of attribute indicator vectors.

By incorporating the constraints discussed above, we can formulate our learning problem. For this purpose, we assume that we have an initial word representation φ (*e.g.* word2vec [98], GloVe [100]) and define our target representation as a projection of it. In the linear case, the transformation is given by:

$$\phi(w) = W\varphi(w) \tag{9}$$

Here, we refer to the W as the transformation matrix[1]. The goal is to learn a matrix W such that it leads to a visually-meaningful word representation that allows effective knowledge transfer from seen classes to unseen classes. Inspired from the structural SVM [119, 120]

[1] In this section, we discuss only the linear case and provide non-linear extensions in Section 4.2.

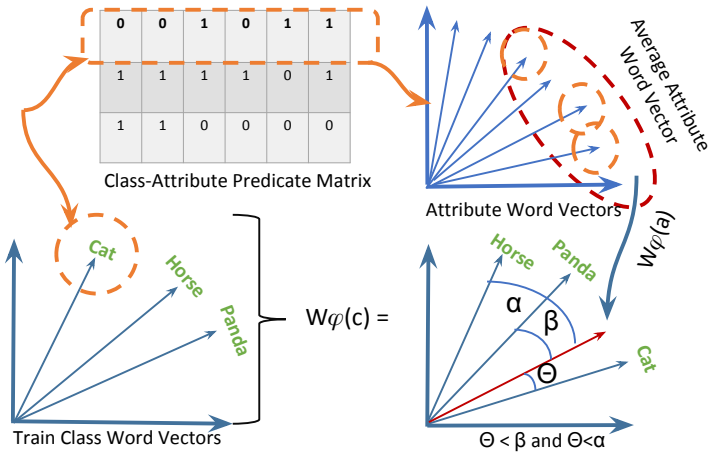


FIGURE 4.1.: Illustration of our approach for learning visually consistent word representations. The goal is to project class and attribute names into a space where the cosine similarity is proportional to the visual similarity between class and attribute names.

formulation, we can define this as a ℓ_2-regularized constrained optimization problem:

$$\begin{aligned} &\min_{W,\xi} \lambda \|W\|_2^2 + \sum_{c \neq c'} \xi_{cc'} \\ &f(c, A_c) \geq f(c, A_{c'}) + \Delta(A_c, A_{c'}) - \xi_{cc'}, \quad \forall c \neq c' \end{aligned} \tag{10}$$

where ξ terms are the slack variables for soft-penalizing unsatisfied similarity constraints, and λ is the regularization weight. Importantly, in this formulation, we implement the principle in Eq. (8) in terms of attribute predicates for the seen classes. Here, we assume that the set A_c of associated attributes are known for each seen (training) class, but unknown for the unseen (zero-shot) classes.

To avoid optimization over non-linear constraints induced by the cosine similarity function, we can write the problem in Eq. (10) as an equivalent unconstrained optimization problem:

$$\min_W \lambda \|W\|_2^2 + \sum_{c \neq c'} \max\left(0, f(c, A_{c'}) - f(c, A_c) + \Delta(A_c, A_{c'})\right) \tag{11}$$

where the function f can be simplified into the following form by plugging in the definition in Eq. (6) and using the linearity in Eq. (6):

$$f(c, A) \;\; = \;\; \sigma\left(W\varphi(c), W\left(\frac{1}{|A|}\sum_{a \in A} \varphi(a)\right)\right) \tag{12}$$

We learn the transformation matrix W by optimizing Eq. (11).

Figure 4.1. illustrates the training stage for our approach. As shown in this figure, the main idea is to project the φ word representations into a new space, where the similarity between a class and an attribute combination in terms of their name vectors is indicative of their visual similarity. Noticeably, the training process is based completely on the class-attribute predicate matrix for the seen classes. Training images are utilized only for building the attribute predictors. Therefore, the training set for the proposed approach can simply be built by creating a table of class and attribute associations.

Zero-shot classification

Once the transformation, *i.e.* the name representation ϕ, is learned, zero-shot classification for unseen classes can be achieved by measuring the similarity $f(c, A)$ between each unseen class name c and a given set of attributes A. On a novel image x, the attribute set A can potentially be obtained via the outputs of a pre-trained attribute classifier. However, in our experiments, instead of using the binary attribute classification outputs, we incorporate the uncertainty $p(a|x)$ by computing a weighted average of attribute name vectors:

$$f_{\text{test}}(c, x) = \sigma\left(W\varphi(c), W\left(\frac{1}{\sum_a p(a|x)}\sum_a p(a|x)\varphi(a)\right)\right) \tag{13}$$

where $f_{\text{test}}(c, x)$ is the discriminant function for zero-shot learning.

Figure 4.2. illustrates our zero-shot classification approach. Given a novel image, we first apply the attribute predictors and compute the weighted average of the attribute name representations. The zero-shot classification is done by comparing the resulting combined attribute name representation with each one of the class name representations in the transformed space. The image can be assigned to the class with the highest cosine similarity.

Learning with image-based training data

In our original formulation in Eq. (11), the training procedure aims to separate classes in the word representation space, proportional to the visual dissimilarity measured by the Hamming distance between class attribute combinations. Here, we show that we can improve the formulation by using image-based training data instead of predicate-based.

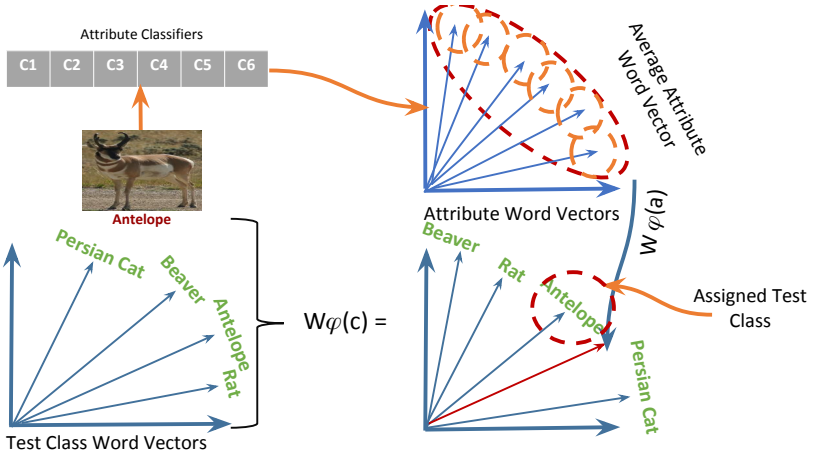


FIGURE 4.2.: Illustration of our zero-shot classification approach (test stage). Prediction depends on the similarity between the class name representation, and the average of attribute name representations, which are parameterized by the transformation matrix W.

Eq. (11) contains of comparison between known classes and their attribute combinations. These combinations are formed according to the binary predicate matrix which contains prior information about class-attribute relationships. However, attribute information for test classes are obtained through attribute classifiers, so they does not contain precise and binary results like predicate matrices. At this point, we can handle Eq. (11) with attribute results which are obtained from validation part of attribute classifier training process.

On the other hand, if we look at our original constraint definition in Eq. (8), it suggests that the similarity of a class to its true attribute combination A_c versus *any* other "negative" attribute combination A' should have a margin value of at least $\Delta(A_c, A')$. However, we lose the "negative" attribute combination information with image-based training data. We can redefine the original constraint definition by protecting the hamming distance part in Eq. (14):

$$f(c, I_a) \geq f(c', I_a) + \Delta(A_c, A'), \qquad \forall c' \neq c \tag{14}$$

where I_a term represents combined attribute vector which is obtained from attribute classifier results. Similarly, Eq. (11) needs to be modified according to the change occurring in Eq. (8):

$$\min_W \lambda \|W\|_2^2 + \sum_{c \neq c'} \max\left(0, f(c', I_a) - f(c, I_a) + \Delta(A_c, A_{c'})\right) \tag{15}$$

Zero Shot Action Recognition

In this part, we provide detailed information about our zero-shot action recognition approach. Zero-shot action recognition has a similar approach with zero-shot object classification, but also contains some differences. These differences can be listed as follows:

- Action recognition datasets do not contain hand-crafted predicate matrices. Therefore, we try to obtain these matrices from the textual data by measuring the cosine similarity between action and object descriptions.
- Attribute annotations are not included in datasets, so we choose ready-to-use ImageNet [69] object classifiers as does the method we need to compare [66].
- Unlike our zero-shot object classification approach, training set consists of different dataset from test set, so all of the action classes in the datasets are handled as zero-shot targets (cross domain experiment).
- Since there is no hand-crafted predicate matrix, we can only try to learn transformation matrix from image-based training data.

In the next section, we evaluate our main approach and the extensions in detail.

4.2. Experiments

In our experiments, we use GloVe [100] and word2vec [98] models to obtain the initial word representations, which is transformed by our method to make it visually, rather than semantically, more consistent. We apply our method on two different zero-shot learning applications, namely zero-shot object classification and zero-shot action recognition. We evaluate our model on four different challenging benchmark datasets, namely AwA [121], aPaY [7], HMDB51 [110] and UCF Sports [111] datasets. We use convolutional neural net features [68, 122] to represent visual information and to train attribute-based classifiers. Our experimental results show that our method yields comparable to or better than the state-of-the-art. In addition, our proposed approach is competitive with its supervised counterparts.

Zero Shot Learning

In this part, we explain our zero-shot object classification experiments on two common datasets (*i.e.* AwA and aPaY). We first briefly introduce word embedding vectors, attribute classifiers and transformation matrices which are used in this experiment. Then, we compare our methods with other zero-shot learning approaches and discuss the experimental results.

Word Embeddings. We use 300-dimensional word embedding vectors generated with GloVe [100] approach that uses Common Crawl Data[2]. These word vectors are publicly available[3]. We generate word vectors for each class and attribute names in the datasets. Some class and attribute names consist of multiple words. In such cases, we generate word vectors for each word and then we take the average of them.

Attribute Classifiers. We use state-of-the-art Convolutional Neural Network (CNN) features to encode images and train attribute classifiers. For each dataset, we utilize the CNN-M2K features [46], where images are resized to 256x256 and mean image subtraction is applied. Five different crops of images and their flipped versions are generated before feeding into the network. Outputs of fc7 layer are used as image representations, resulting in 2,048 dimensional feature vectors per image.

Following Farhadi *et al*. [7], we obtain our attribute classifiers by training ℓ_2-regularized squared-hinge-loss linear SVMs. Parameter selection is done using 10-fold cross validation over the training set. We map the attribute prediction scores to posterior probability via Platt scaling.

Attribute based classifiers require class-based attribute labels during training. AwA dataset provides class-level attribute annotations, but aPaY only provides image-level attribute annotations. Class-level attribute annotations for these datasets are produced using image-level attribute annotations, such that an attribute is marked as positive for a class if any of images in the class contains the related attribute.

Word Representation Learning. We optimize transformation matrix using the two layered feed-forward network. Parameter selection (*i.e.* number of hidden unit) is done using 2-fold cross validation on AwA and aPaY attribute datasets. Network details are as follows:

[2]http://commoncrawl.org/the-data/
[3]http://nlp.stanford.edu/projects/glove/

TABLE 4..1: Attribute classification performances of different methods. Mean AUC has been used as the metric of comparison.

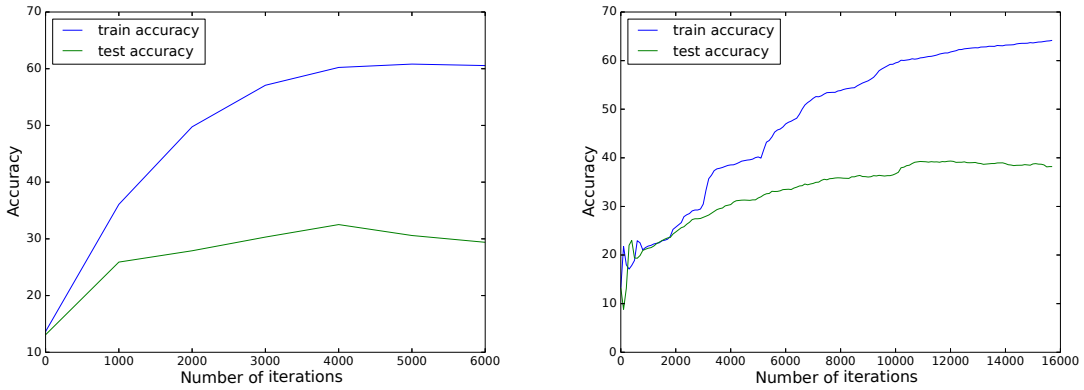
Dataset	**Features**	HAT[46]	DAP[22]	Our
AwA	Shallow	71.16	72.8	-
	Deep	78.64	78.00	80.56
aPaY	Shallow	70.91	-	-
	Deep	80.73	-	84.91

- Number of hidden unit is selected from [100, 200, 300, 400, 500] values.
- Adam [123] is used for stochastic optimization, and learning rate value is selected as *1e-4*.
- *tanh* function is used as the activation function in the first hidden layer.
- *sigmoid* function is used as the activation function in the second hidden layer.
- Implementation is done with the help of TensorFlow[124].

Results. In our experiments, we first evaluate the performance of our attribute classifiers, since their performance is likely to have an influence on knowledge transfer. We measure the success of the attribute classifiers using mean AUC and the results are given in Table 4..1. We obtain 80.56% mean attribute prediction performance on AwA dataset and 84.91% on aPaY dataset. According to these results, we observe that our attribute classifiers based on deep features on AwA and aPaY datasets are quite usable.

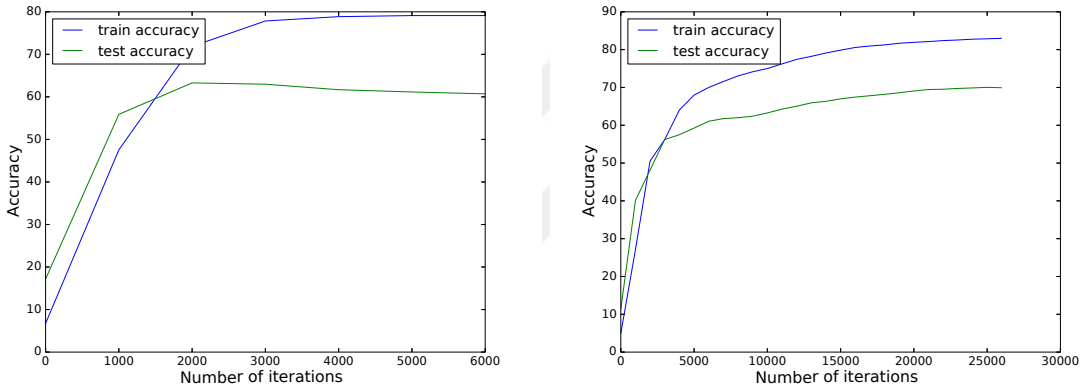
Table 4..2 presents the experimental results obtained via different options of our algorithm. We report normalized top-1 per-class averaged accuracy for zero-shot learning. Here, baseline method is created without the use of any learned transformation matrix. Moreover, PBW (Predicate Based W) represents our proposed approach that learns a transformation matrix with using predicate matrix, whereas IBW (Image Based W) represents learning transformation matrix with using training images. According to the results presented in Table 4..2, learning transformation matrix is very useful and using image based training data yields the most successful results on the aPaY and AwA datasets.

We further investigate whether the number of epoch on the transformation matrix learning stage has an impact on the classification. As it can be seen on Figure 4.3. and 4.4., increasing the number of epoch makes a positive impact to the accuracy in general, but this impact remains stable after a while.



(a) PBW (b) IBW

FIGURE 4.3.: Top-1 per-class averaged accuracy results during training steps (aPaY Dataset)

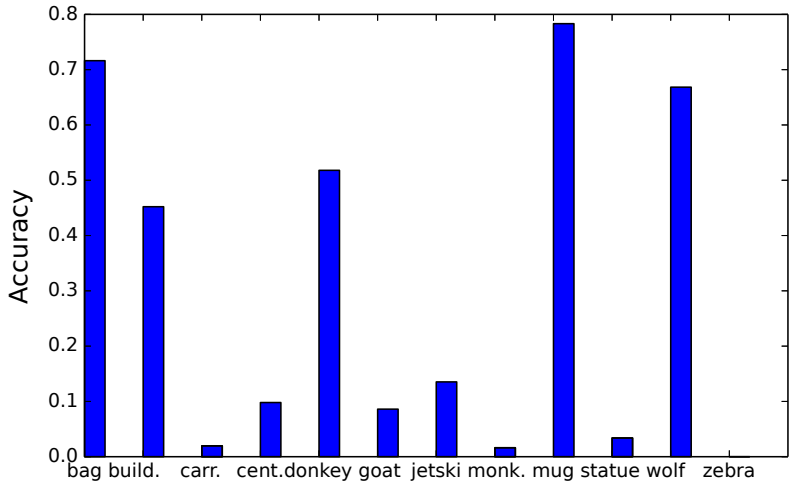


(a) PBW (b) IBW

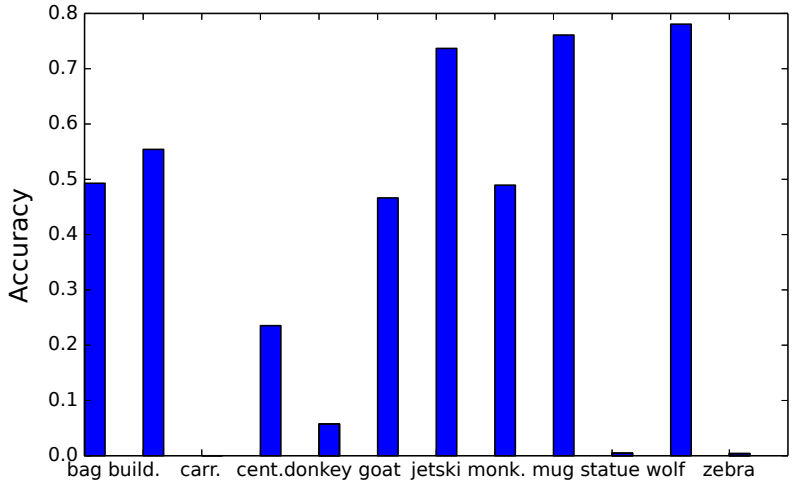
FIGURE 4.4.: Top-1 per-class averaged accuracy results during training steps (AwA Dataset)

Figure 4.5. and 4.6. show the class-wise prediction accuracies of our methods. As can be seen that the results of our methods are more stable and successful in AwA dataset. In aPaY dataset, we observe that, for classes *carriage*, *centaur*, *statue* and *wolf*, our methods produce worse results than other classes. There are two main reasons for this situation: (i) training classes are very different from these test classes, (ii) there are visually similar classes to these classes in the test set. Moreover, similar information and arguments are also shown on confusion matrices in Figure 4.7. and 4.8..

We then compare our approach to various unsupervised [48, 49, 89, 90, 96, 125, 126] and supervised [22, 46, 47] counterparts presented in the literature. The results are shown on Table 4..3. Here, supervision corresponds to the information needed during test for zero-shot

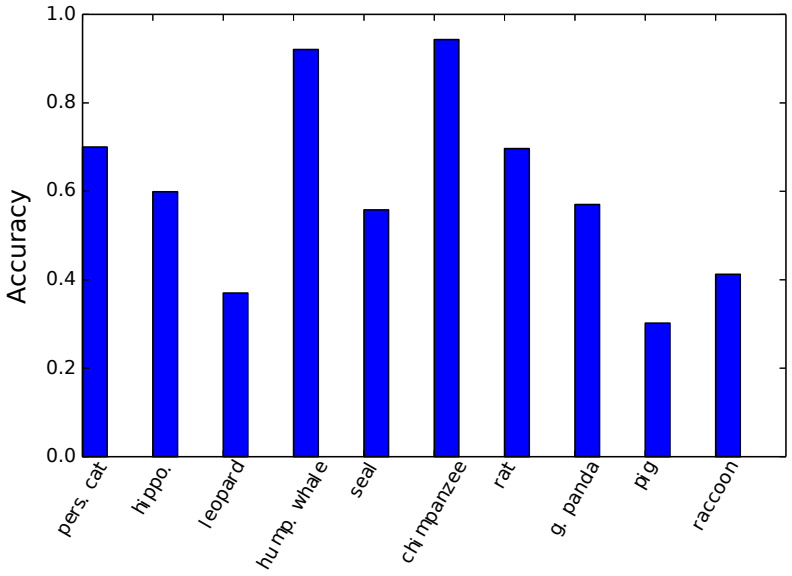


(a) PBW

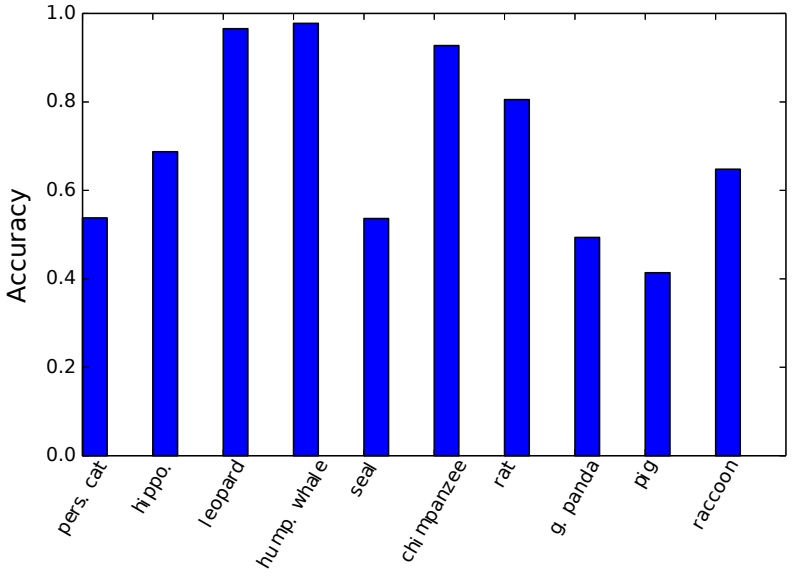


(b) IBW

FIGURE 4.5.: Class-wise prediction accuracies of our methods (aPaY Dataset)

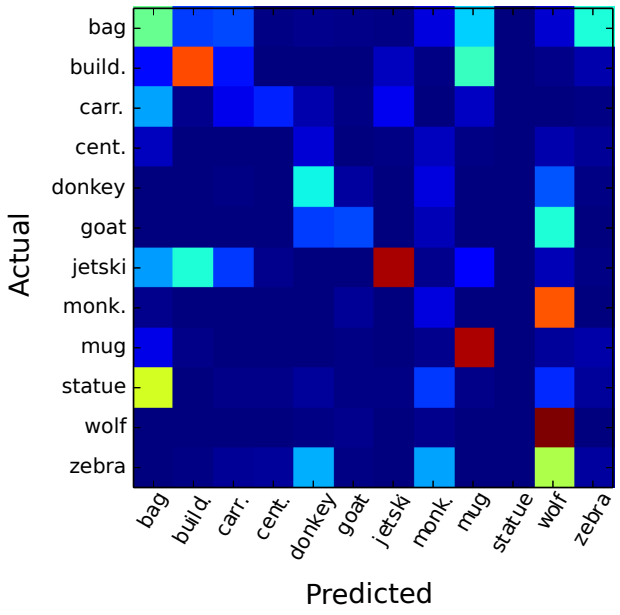


(a) PBW

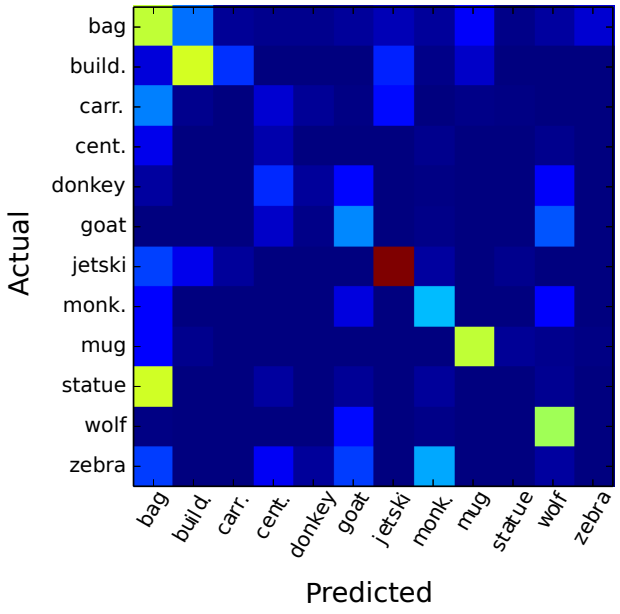


(b) IBW

FIGURE 4.6.: Class-wise prediction accuracies of our methods (AwA Dataset)

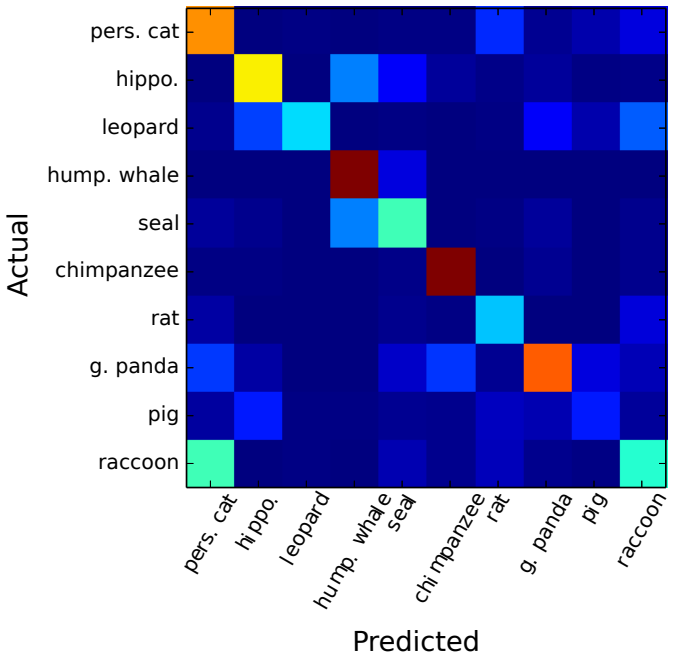


(a) PBW

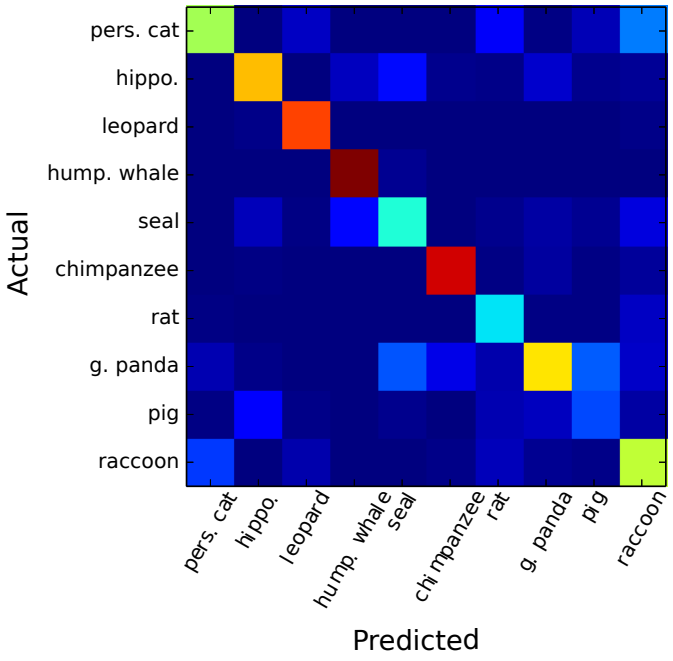


(b) IBW

FIGURE 4.7.: Confusion matrices of our methods (aPaY Dataset)



(a) PBW



(b) IBW

FIGURE 4.8.: Confusion matrices of our methods (AwA Dataset)

TABLE 4..2: Zero-shot learning results of our methods. We report normalized top-1 per-class averaged accuracy.

Method	AwA	aPaY
Baseline	10.2	16.0
PBW	60.71	29.38
IBW	**69.92**	**38.18**

TABLE 4..3: Zero-shot learning results of various unsupervised and supervised methods on two different datasets.

test supervision	method	AwA	aPaY
unsupervised	ALE[49]	59.8	33.3
	LatEm[96]	62.9	-
	DeViSE[90]	44.5	25.5
	ConSE[89]	46.1	22.0
	Text2Visual[48, 125]	55.3	30.2
	CAAP[126]	67.5	37.0
	Our method	**69.92**	**38.18**
supervised	DAP[22]	54.0	28.5
	ENS[47]	57.4	31.7
	HAT[46]	63.1	38.3

learning; the supervised methods require additional data about the unseen classes such as attribute-class predicate matrices, whereas unsupervised methods do not require any inputs. Hence, supervised methods have a big advantage in this comparison because they know attribute signatures of test classes. On the other hand, unsupervised methods infers attribute signatures using the training set only (or in the presence of additional auxiliary data as in the case of ALE [49]). Note that, there are several other methods in the literature reporting performance for zero-shot learning on these datasets, however, since they operate on other sets of low-level visual features, the results are not directly comparable. Therefore, for the sake of fair comparison, we only compare to those methods that use the same set of image representations.

When we review the results on Table 4..3, we observe that on AwA and aPaY datasets, our method yields state-of-the-art classification performance according to its unsupervised counterparts. Our method also produces similar performance with respect to supervised methods, producing only 0.12% less accuracy than the HAT [46] method on aPaY dataset. It achieves the best results with a large margin in AwA dataset.

FIGURE 4.9.: Top-5 highest score images for each class in the AwA dataset using deep features. These results are obtained from DAP [22] method.

Figure 4.9., 4.10. and 4.11. illustrate qualitative examples of the results of our transformation matrix learning approach. We compare the top-5 scoring images of our method to that of DAP[22] in this visual analysis. The results demonstrate that in AwA dataset, our image based method produces significantly better results in some classes. Moreover, based on these results, we can say that each zero-shot learning framework encodes a different aspect of the dataset.

Zero Shot Action Recognition

In this part, we provide detailed information about our zero-shot action recognition experiment on two action recognition datasets (*i.e.* UCF-Sport and HMDB51).

Rest of the this section is as follows. We first briefly introduce word embedding vectors, object classifiers and transformation matrix which is used in this experiment. Then, we

FIGURE 4.10.: Top-5 highest score images for each class in the AwA dataset using deep features. These results are obtained from our method using predicate based transformation matrix.

compare our methods with other zero-shot action recognition approaches and discuss the experimental results.

Word Embeddings. We use 500-dimensional word embedding vectors generated with skip-gram model of word2vec [98] approach which uses YFCC100M [127] dataset. YFCC100M dataset contains metadata tags of about 100M Flickr images and the word vectors obtained from YFCC100M are publicly available[4]. We use these word vectors directly in our experiments as it used in the approach which we want to compare.

Object Classifiers. We use state-of-the-art object classification scores to encode actions and learn proper transformation matrices. For each dataset, we utilize AlexNet[68] classification scores, where every 10th frame is sampled. Each sampled frames are represented with total 15,293 ImageNet object categories which have more than 100 examples. After that, average

[4]https://staff.fnwi.uva.nl/m.jain/projects/Objects2action.html

FIGURE 4.11.: Top-5 highest score images for each class in the AwA dataset using deep features. These results are obtained from our method using image based transformation matrix.

pooling is applied on sampled frames for each action videos, so each video is represented with 1xd[d=15,293] dimensional vectors.

We used the action scores which are published by the approach we want to compare for a fair comparison. These scores are again publicly available[5].

Word Representation Learning. We implement our model wih using TensorFlow[124]. Parameter selection (*i.e.* number of matrix dimension) is done using 2-fold cross validation on UCF-101 dataset. Model details are as follows:

- Number of hidden unit is selected from [100, 200, 300, 400, 500] values.
- Adam [123] is used for stochastic optimization, and learning rate value is selected as *1e-4*.

[5]https://staff.fnwi.uva.nl/m.jain/projects/Objects2action.html

TABLE 4..4: Zero-shot action recognition results on two different dataset.

Method	UCF-Sport	HMDB51
DAP[22]	11.67	01.94
objects2action[66]	23.92	08.60
Our Method	**26.67**	**08.78**

- Only image-based loss is used because datasets do not contain hand-crafted predicate matrices.

Since we do not have any training data, we train our transformation matrix with a different dataset(*i.e.* UCF-101). However, there are common action classes in both the training and test sets and these are eliminated from training dataset for a correct zero-shot learning setting. Some of these common classes are as follows: *Diving, HorseRiding, PushUps, PullUps, Punch, Biking, Fencing*.

Results. In the experiments, we report normalized top-1 per-class averaged accuracy for zero-shot action recognition. We compare our approach with objects2action [66] and DAP [22] methods. The results are shown on Table 4..4. When we review the results on Table 4..4, we observe that on UCF-Sport and HMDB51 datasets, our method yields state-of-the-art action classification performance according to the its counterparts. These results show that transformation matrix can carry semantic information not only between training and test sets but also datasets. In addition, the HMDB dataset contains a large number of action classes according to other data sets and we achieved more successful results than other studies. Therefore, we can say that our approach is more suited for large-scale experiments than others.

4.3. Conclusions

An important limitation of the existing methods for zero-shot learning is their dependency on the attribute signatures of the unseen classes or auxiliary textual data. To eliminate this dependency, in this work, we utilize attributes as an intermediate representation, in an unsupervised way for the unseen classes. To this end, we learn a visually consistent word representation such that the similarities between class and attribute name representations depict the visual similarity, and use this learned representation to transfer knowledge from seen to unseen classes. Our proposed zero-shot learning method is easily scalable to work with

any unseen class without requiring manually defined attribute-class annotations or any type of auxiliary data.

Experimental results on several benchmark datasets demonstrate the efficiency of our approach, establishing the state-of-the-art amongst the unsupervised methods, at the same time yielding comparable performance to its supervised counterparts.

5. CONCLUSION

In this thesis, we prepared novel methods for top-down saliency estimation and unsupervised zero-shot object recognition problems. The developed methods are important in that they show the advantages of using the attribute information. Attributes can describe visual appearance, functional affordance or human-understandable aspects of objects, and this thesis tries to exploit mentioned features of the attributes to improve defined problems.

In the visual saliency estimation problem, researchers try to predict where humans look at images through different computational models. The method we have developed to improve this problem is based on using attribute classifier results to encode visual features of images. Experimental results show that attribute based classifiers generate more reliable and compact results than SIFT descriptors.

Zero-shot learning tries to classify images of classes that are not seen before by using limited prior knowledge. In the zero-shot learning problem, we learn a visually consistent word representation such that the similarities between class and attribute name representations depict the visual similarity and lead to accurate classifications. Experimental results obtained from several benchmark datasets demonstrate the efficiency of our approach, establishing the state-of-the-art amongst the unsupervised methods, at the same time yielding comparable performance to its supervised counterparts.

5.1. Main Contributions

Top-Down Visual Saliency Estimation

- We develop a novel approach to improve top-down saliency estimation problem. According to the proposed approach, model uses middle-level features to encode visual information in images and their local parts. In this context, model uses results of attribute based classifiers as feature vectors.
- We have demonstrated that the attribute based classifiers generate more reliable and compact results than SIFT descriptors. Moreover, these classifiers produce these more accurate results with smaller feature dimensionality. This situation shows that the attribute based classifiers encode visual information better than SIFT descriptors because

they expose high-level object properties by combining scores of different kind of attributes and by additional information.

Zero-Shot Object Classification

- We develop a novel method for learning a visually consistent word representation to improve zero-shot learning problem. The proposed method uses class-attribute relationships to learn semantic transformation matrix which transfers semantic information from seen training classes to the unseen test classes.
- This method is utilizing attributes as an intermediate representation in an unsupervised way for the unseen classes because the greatest disadvantage of most of the existing methods for zero-shot learning is their dependency on the attribute signatures of the unseen classes or auxiliary textual data.
- This zero-shot learning method is easily scalable to work with any unseen class without requiring manually defined attribute-class annotations or any type of auxiliary data.

5.2. Future Work

Top-Down Visual Saliency Estimation

- Attribute based classifiers can be used more efficiently. These classifiers are trained on aPascal dataset and it contains 64 different attributes that are belong to 20 different object classes. Some of these object classes are irrelevant to our target objects, so dataset contains irrelevant attributes for our purpose. In our approach, we use all attribute results regardless of the type of the target object, but choosing appropriate and relevant ones among attributes might affect success rate positively. For example, if we try to learn a model for person class, we can use only person related attributes.
- We can give different impact weights to the attributes according to the target object class.
- Using objectness maps [128–131] or superpixel based representations might increase our accuracy [38]. The main advantage of superpixel based approaches is that they can encode object boundary information more successfully than patch-based approaches.

Zero-Shot Object Classification

- Exploration of better formalizations of distributional word vector representations produces more successful results for the problem of zero-shot learning.
- Multiple transformation matrices can be learned instead of learning only one transformation matrix. Semantic information carrying capacity of a transformation matrix might not be enough, so different hidden semantic information can be held in different transformation matrices.
- Attribute based classifiers can be learned from an end-to-end deep model directly. More successful attribute based classifiers can lead to the more successful results.
- Unlike other method, our method uses only textual data during training process, so the success of the method can be improved by using additional textual data.

REFERENCES

[1] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, pages 438–451. Springer, **2010**.

[2] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2973–2980. IEEE, **2012**.

[3] Rogerio Feris, Behjat Siddiquie, Yun Zhai, James Petterson, Lisa Brown, and Sharath Pankanti. Attribute-based vehicle search in crowded surveillance videos. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 18. ACM, **2011**.

[4] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *European Conference on Computer Vision*, pages 354–368. Springer, **2012**.

[5] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *Computer Vision–ECCV 2012*, pages 609–623. Springer, **2012**.

[6] Zhiyuan Shi, Timothy M Hospedales, and Tao Xiang. Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4184–4193. **2015**.

[7] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, **2009**.

[8] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372. IEEE, **2009**.

[9] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE, **2011**.

[10] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, **2012**.

[11] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, **2011**.

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, pages 530–543. Springer, **2012**.

[13] Huizhong Chen, Andrew C Gallagher, and Bernd Girod. What's in a name? first names as facial attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3373. **2013**.

[14] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, **2010**.

[15] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, **2011**.

[16] Shuo Wang, Jungseock Joo, Yizhou Wang, and Song-Chun Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3111–3118. **2013**.

[17] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *2011 International Conference on Computer Vision*, pages 1543–1550. IEEE, **2011**.

[18] Gaurav Sharma and Frederic Jurie. Learning discriminative spatial representation for image classification. In *BMVC*, pages 1–11. BMVA Press, **2011**.

[19] Lucy Liang and Kristen Grauman. Beyond comparing image pairs: Setwise active learning for relative attributes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 208–215. **2014**.

[20] Adriana Ivanova Kovashka. *Interactive image search with attributes*. Ph.D. thesis, **2015**.

[21] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. **2009**.

[22] C.H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, **2014**.

[23] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model the bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28:802–817, **2006**.

[24] Thomas Mauthner, Horst Possegger, Georg Waltner, and Horst Bischof. Encoding based saliency detection for videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2494–2502. **2015**.

[25] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2049–2056. IEEE, **2006**.

[26] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, **2006**.

[27] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages II–37. IEEE, **2004**.

[28] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *IEEE transactions on multimedia*, 7(5):907–919, **2005**.

[29] Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1139–1146. **2013**.

[30] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 733–740. IEEE, **2012**.

[31] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173. **2013**.

[32] Erkut Erdem and Aykut Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11–11, **2013**.

[33] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821. **2014**.

[34] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890. **2014**.

[35] Ruth Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39(19):3157–3163, **1999**.

[36] Ruth Rosenholtz. Search asymmetries? what search asymmetries? *Perception & Psychophysics*, 63(3):476–489, **2001**.

[37] Jimei Yang and Ming-Hsuan Yang. Top-down visual saliency via joint crf and dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2296–2303. IEEE, **2012**.

[38] Aysun Kocak, Kemal Cizmeciler, Aykut Erdem, and Erkut Erdem. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*. **2014**.

[39] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, pages 241–248. **2008**.

[40] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, **2012**.

[41] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. **2012**.

[42] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, **2006**.

[43] Nazar Khan and Marshall F Tappen. Discriminative dictionary learning with spatial priors. In *2013 IEEE International Conference on Image Processing*, pages 166–170. IEEE, **2013**.

[44] Marcin Marszałek and Cordelia Schmid. Accurate object recognition with shape masks. *International journal of computer vision*, 97(2):191–209, **2012**.

[45] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289. **2001**.

[46] Ziad Al-Halah and Rainer Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 837–843. IEEE, **2015**.

[47] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1641–1648. IEEE, **2011**.

[48] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero shot learning using purely textual descriptions. In *ICCV*. **2013**.

[49] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE Computer Society, **2015**.

[50] Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*. **2015**.

[51] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, **1999**.

[52] David G Lowe. Local feature view clustering for 3d object recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–682. IEEE, **2001**.

[53] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, **2004**.

[54] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pages 433–440. **2007**.

[55] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1777–1784. IEEE, **2011**.

[56] Yu Su and Frédéric Jurie. Learning compact visual attributes for large-scale image classification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 51–60. Springer, **2012**.

[57] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, pages 1–14. Springer, **2012**.

[58] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *Computer Vision–ECCV 2008*, pages 340–353. Springer, **2008**.

[59] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 801–808. IEEE, **2011**.

[60] Jingen Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*. **2011**.

[61] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*. **2011**.

[62] Gang Wang and David Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *2009 IEEE 12th International Conference on Computer Vision*, pages 537–544. IEEE, **2009**.

[63] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. *arXiv preprint arXiv:1607.08085*, **2016**.

[64] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *arXiv preprint arXiv:1607.02104*, **2016**.

[65] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 819–826. IEEE, **2013**.

[66] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4588–4596. **2015**.

[67] Amir Sadovnik, Andrew Gallagher, Devi Parikh, and Tsuhan Chen. Spoken attributes: Mixing binary and relative attributes to say the right thing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2160–2167. **2013**.

[68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. **2012**.

[69] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, **2009**.

[70] Yann LeCun, LD Jackel, Leon Bottou, A Brunot, Corinna Cortes, JS Denker, Harris Drucker, I Guyon, UA Muller, Eduard Sackinger, et al. Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks*, volume 60, pages 53–60. **1995**.

[71] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, **1998**.

[72] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, **2013**.

[73] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. **2015**.

[74] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, **2013**.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, **2015**.

[76] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, **2014**.

[77] Hisham Cholakkal, Jubin Johnson, and Deepu Rajan. Backtracking scspm image classifier for weakly supervised top-down saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5278–5287. **2016**.

[78] Hisham Cholakkal, Jubin Johnson, and Deepu Rajan. Weakly supervised top-down salient object detection. *arXiv preprint arXiv:1611.05345*, **2016**.

[79] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, **2011**.

[80] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 478–485. IEEE, **2012**.

[81] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1976–1983. **2013**.

[82] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 153–160. **2013**.

[83] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision–ECCV 2014*, pages 48–64. Springer, **2014**.

[84] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, pages 3464–3472. **2014**.

[85] Bernardino Romera-Paredes and PHS Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2152–2161. **2015**.

[86] Svetlana Kordumova, Thomas Mensink, and Cees GM Snoek. Pooling objects for recognizing scenes without examples. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 143–150. ACM, **2016**.

[87] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. **2011**.

[88] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, **2001**.

[89] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, **2013**.

[90] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129. **2013**.

[91] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418. **2009**.

[92] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943. **2013**.

[93] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, **2010**.

[94] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 18(5):401–409, **1969**.

[95] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. *arXiv preprint arXiv:1603.08754*, **2016**.

[96] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. *arXiv preprint arXiv:1603.08895*, **2016**.

[97] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. *arXiv preprint arXiv:1509.04767*, **2015**.

[98] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. **2013**.

[99] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. **2013**.

[100] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, **2014**.

[101] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, **1995**.

[102] Christiane Fellbaum. *WordNet*. Wiley Online Library, **1998**.

[103] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics, **2004**.

[104] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, **1990**.

[105] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, **2013**.

[106] Daniel N Osherson, Joshua Stern, Ormond Wilkie, Michael Stob, and Edward E Smith. Default probability. *Cognitive Science*, 15(2):251–269, **1991**.

[107] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5. **2006**.

[108] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, **2010**.

[109] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, **2012**.

[110] HA Jhuang, HA Garrote, EA Poggio, TA Serre, and T Hmdb. A large video database for human motion recognition. In *Proc. of IEEE International Conference on Computer Vision*. **2011**.

[111] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, **2014**.

[112] Marcin Marszalek and Cordelia Schmid. Accurate object localization with shape masks. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, **2007**.

[113] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431, **2006**.

[114] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., **1967**.

[115] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, **1995**.

[116] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, **2005**.

[117] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, **1986**.

[118] Gernot Hoffmann. Cielab color space. *Wikipedia, the free encyclopedia. mht*, **2003**.

[119] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484, **2005**.

[120] Ben Taskar Carlos Guestrin Daphne Roller. Max-margin markov networks. *Advances in neural information processing systems*, 16:25, **2004**.

[121] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, **2009**.

[122] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, **2014**.

[123] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, **2014**.

[124] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org*, 1, **2015**.

[125] Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28–52, **2010**.

[126] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning.

[127] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, **2015**.

[128] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, **2010**.

[129] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3286–3293. **2014**.

[130] Ian Endres and Derek Hoiem. Category independent object proposals. In *European Conference on Computer Vision*, pages 575–588. Springer, **2010**.

[131] Pekka Rantalankila, Juho Kannala, and Esa Rahtu. Generating object segmentation proposals using global and local search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2417–2424. **2014**.

CURRICULUM VITAE

Credentials

Name,Surname : Berkan DEMIREL
Place of Birth : ANKARA
Marital Status : Single
E-mail : demirelberkan@gmail.com
Address : Computer Engineering Department of Hacettepe University
Beytepe,ANKARA

Education

MS : Computer Engineering, Hacettepe University, Ankara, TURKEY, (ongoing)
BSc : Computer Engineering, Hacettepe University, Ankara, TURKEY, 2014

Foreign Languages

English

Work Experience

R&D Engineer, HAVELSAN Inc., 2015-?
Software Engineer, MilSOFT Inc., 2013-2015
Intern, HAVELSAN Inc., 2013
Intern, MilSOFT Inc., 2012

Areas of Experiences

Image Processing, Computer Vision, Remote Sensing

Projects and Budgets

——

Publications

”Visual Saliency Estimation via Attribute Based Classifiers and Conditional Random Field”,
22nd Signal Processing and Communications Applications Conference (SIU), May 2016,
Berkan Demirel, Ramazan Gokberk Cinbis, Nazli Ikizler Cinbis
”Hyperspectral Image Segmentation Based on Spatial Model”,
22nd Signal Processing and Communications Applications Conference (SIU), May 2016,
Berkan Demirel, Omer Ozdil, Yunus Emre Esin

Oral and Poster Presentations

SIU 2016, Zonguldak

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF Computer Engineering

Date: 6/12/2016

Thesis Title / Topic: ATTRIBUTE BASED CLASSIFIERS FOR IMAGE UNDERSTANDING

According to the originality report obtained by ~~myself~~/my thesis advisor by using the Turnitin plagiarism detection software and by applying the filtering options stated below on 6/12/2016 for the total of 80 pages including the a) Title Page, b) Introduction, c) Main Chapters, and d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 8 %.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes ~~excluded~~ included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Science and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

Name Surname: Berkan Demirel
Student No: N14125263
Department: Computer Engineering
Program: Masters
Status: ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Assist Prof Nazlı İkizler Cinbiş
(Title, Name Surname, Signature)

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
Bilgisayar Mühendisliği **ANABİLİM DALI BAŞKANLIĞI'NA**

Tarih: 6/12/2016

Tez Başlığı / Konusu: Görüntü Anlamlandırmak için Nitelik Tabanlı Sınıflandırıcılar

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve d) Sonuç kısımlarından oluşan toplam 80 sayfalık kısmına ilişkin, 6/12/2016 tarihinde ~~şahsım~~/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 8 'tür.

Uygulanan filtrelemeler:
1- Kaynakça hariç
2- Alıntılar ~~hariç~~/dâhil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

Adı Soyadı:	Berkan Demirel
Öğrenci No:	N14125263
Anabilim Dalı:	Bilgisayar Mühendisliği
Programı:	Yüksek Lisans
Statüsü:	☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Yrd. Doç. Dr. Nazlı İkizler Cinbiş
(Unvan, Ad Soyad, İmza)