Advances in Computer Vision and Pattern Recognition

Rogerio Schmidt Feris
Christoph Lampert
Devi Parikh *Editors*

Visual Attributes

Springer

Advances in Computer Vision and Pattern Recognition

More information about this series at http://www.springer.com/series/4205

Rogerio Schmidt Feris · Christoph Lampert
Devi Parikh
Editors

Visual Attributes

Springer

Editors
Rogerio Schmidt Feris
IBM T.J. Watson Research Center
Yorktown Heights, NY
USA

Devi Parikh
Georgia Tech
Atlanta, GA
USA

Christoph Lampert
Computer Vision and Machine Learning
IST Austria
Klosterneuburg
Austria

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-3-319-50075-1 ISBN 978-3-319-50077-5 (eBook)
DOI 10.1007/978-3-319-50077-5

Library of Congress Control Number: 2016958717

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Visual attributes are generally defined as mid-level semantic visual concepts or properties that are shared across categories, e.g., furry, striped, metallic, young. They have recently gained significant popularity in computer vision, finding applications in zero-shot classification (where a machine can recognize a concept even without having seen it before), image ranking and retrieval, fine-grained categorization, human–machine interaction, and many others.

This book provides an overview of and summarizes recent advances in machine learning and computer vision related to visual attributes, while exploring the intersection with other disciplines such as computational linguistics and human–machine interaction. It contains a collection of chapters written by world-renowned scientists, covering theoretical aspects of visual attribute learning as well as practical computer vision applications.

We would like to express our sincere gratitude to all chapter contributors for their dedication and high-quality work, as well as to Simon Rees and Wayne Wheeler from Springer for their support and help throughout the book's preparation.

Yorktown Heights, NY, USA — Rogerio Schmidt Feris
Vienna, Austria — Christoph Lampert
Atlanta, GA, USA — Devi Parikh
September 2016

Contents

Chapter 1
Introduction to Visual Attributes

Rogerio Schmidt Feris, Christoph Lampert and Devi Parikh

Visual recognition has significantly advanced in recent years, particularly through the widespread adoption of deep convolutional neural networks [22, 28] as the main tool for solving computer vision problems. The recognition accuracy recently obtained in standard benchmark datasets, such as Imagenet [7], has even surpassed human-level performance [15].

The fuel to power up these neural network models is training data. In fact, current methods often require at least thousands of manually annotated training examples for learning robust classifiers for new categories. While it is easy to obtain a large number of example images for common categories, such as images of vehicles or dogs, it is not straightforward to obtain annotated training sets for other infrequent categories, such as a particular vehicle model or a specific dog breed. There are tens of thousands of basic categories in the world (and significantly more subordinate categories) [3]. For many of them, only a few or *no examples at all* are available.

Zero-data or zero-shot classification refers to the problem of recognizing categories for which no training examples are available [26, 30]. This problem happens in many practical settings. As an example, for the task of predicting concrete nouns from neural imaging data [30], many nouns may not have corresponding neural image examples because of the costly label acquisition process. In the visual surveillance domain, while conducting a criminal investigation, the police may have only

R.S. Feris (✉)
IBM T. J. Watson Research Center, New York, USA
e-mail: rsferis@us.ibm.com

C. Lampert
Institute of Science and Technology Austria, Klosterneuburg, Austria
e-mail: chl@ist.ac.at

D. Parikh
Georgia Tech, Atlanta, GA, USA
e-mail: parikh@gatech.edu

R.S. Feris et al. (eds.), *Visual Attributes*, Advances in Computer Vision and Pattern Recognition, DOI 10.1007/978-3-319-50077-5_1

eyewitness descriptions available for searching a targeted suspect, instead of example images [13, 40]. Many *fine-grained* visual categorization tasks have classes for which only a few or no training images exist. For instance, the ImageNet dataset has 30 mushroom synsets, each with 1000 images, whereas there are more than ten thousand mushroom species found in nature. The zero-shot classification problem is also common in other fields. In large vocabulary speech recognition systems, it is infeasible to acquire training samples for each word. Recommender systems face issues when new apps are released without any user ratings (also known as the cold-start problem [35]).

Visual attributes, which are generally defined as mid-level semantic properties that are shared across categories (e.g., furry, yellow, four-legged), provide an effective way of solving the zero-shot classification problem. As initially demonstrated by Lampert et al. [25, 26], a novel unseen category with an associated description based on semantic attributes (either provided by experts or mined from language sources, such as Wikipedia [33, 34]) can be recognized by leveraging visual attribute classifiers, which can be learned using existing training data from known categories. This process is aligned with human capabilities of identifying objects only based on descriptions. For example, when given a sentence like "large gray animals with long trunks," we can reliably identify elephants [26]. Currently, the highest-performing methods for zero-shot learning rely on visual attributes, often in connection with other forms of semantic embedding such as distributional word vector representations [1, 2, 14, 33].

Visual attributes are both semantic (human-understandable) and visual (machine-detectable). In addition to zero-shot learning, they have proven effective in various other applications. As a communication channel between humans and machines, attributes have been used for interactive recognition of fine-grained categories [4], active learning [21], and image search with humans in the loop [20]. Attributes discretize a high dimensional feature space into a simple and readily interpretable representation that can be used to explain machine decisions to humans [16] and predict user annoyance [5]. Conversely, humans can provide rationales to machines as a stronger form of supervision through visual attributes [10]. Along this direction, attributes can serve as a form of privileged information [36] for improving recognition, especially when only a few training examples are available.

Another area in which attributes have recently played a major role is visual analysis of people. In the visual surveillance domain, state-of-the-art person reidentification systems [27, 37, 39] benefit from human attributes as features for improving matching of people across cameras. The extraction of face and clothing attributes enable search for suspects or missing people based on their physical description [13, 40]. In e-commerce applications, attributes are very effective in improving clothing retrieval [17] and fashion recommendation [29]. It has also been shown that facial attribute prediction is helpful as an auxiliary task for improving face detection [42] and face alignment [43]. Methods for image ranking and retrieval also benefit from attributes as a compact and semantic image representation [11, 23, 38].

Other applications of visual attributes include describing unfamiliar objects [12], scene analysis [32], material classification [6], and image virality prediction [8].

Beyond semantics, attributes have been used for understanding and predicting the memorability and aesthetics of photographs [9, 18, 19]. Finally, attributes have been recently used for image editing (e.g., allowing users to adjust the attributes of a scene to be "snowy" or "sunset") [24] and for conditional image generation in the context of generative adversarial networks [41].

This book's goal is to summarize the main ideas related to visual attributes that were proposed in the past few years, and to cover recent research efforts related to this emerging area in an accessible manner to a wider research community. Next, we provide an overview of the chapters of the book, which comprise both theoretical aspects of attribute learning and practical applications.

1.1 Overview of the Chapters

Part I: Attribute-Based Recognition

The first part of the book covers attribute-based methods for *recognition of unseen classes* for which training examples are unavailable (i.e., zero-shot classification), *recognition of seen classes*, where attributes are used as privileged information during the training stage, and methods for *multitask attribute learning*.

Chapter 2, by Bernardino Romera-Paredes and Philip H.S. Torr, introduces the problem of zero-shot learning and proposes a general framework that models the relationships between features, attributes, and classes, so the knowledge learned at the training stage can be transferred to the inference stage. The method is easily implemented: one line of code for training and another for inference; yet, it achieves impressive results on standard benchmark datasets.

In Chap. 3, Viktoriia Sharmanska and Novi Quadrianto consider the problem of visual recognition of categories when their attributes are used as privileged information during training time. In particular, they address whether attributes are still useful privileged data when modern deep convolutional features are used for visual classification. Their analysis shows that the answer to this question depends on the classification task's complexity.

In Chap. 4, Chao-Yeh Chen, Dinesh Jayaraman, Fei Sha, and Kristen Grauman address the problem of multitask attribute learning, exploring when and to what extent sharing is useful for attribute learning. They introduce the idea of selective sharing during multitask learning of attributes, using semantic knowledge to decide what to share and what not to share during learning.

Part II: Relative Attributes and Their Application to Image Search

The second part of the book introduces the concept of relative attributes [31], which consists of measuring the relative strength of properties (for example, "bears are furrier than giraffes") instead of simply determining whether they are present

or not, and demonstrates the effectiveness of modeling relative attributes in image search applications.

In Chap. 5, Adriana Kovashka and Kristen Grauman show how semantic attributes can be effectively used for interactive image search with user feedback based on relative attribute comparisons. They present a system called "WhittleSearch," which can answer queries such as "show me shoes like these, but more formal." This chapter also covers techniques for actively selecting images for feedback and adapting attribute models for personalized user queries.

Chapter 6, by Aron Yu and Kristen Grauman, addresses the problem of fine-grained visual comparisons with attributes, which is valuable for sophisticated image search systems that may need to distinguish subtle properties between highly similar images. They develop computational models based on *local learning* for fine-grained visual comparisons, where a predictive model is trained on the fly using only the data most relevant to a given input. They also address the problem of determining when an image pair is indistinguishable in terms of a given attribute.

In Chap. 7, Fanyi Xiao and Yong Jae Lee introduce a weakly supervised method for automatically discovering the spatial extent of relative attributes in images. This is achieved by mining a set of local, transitive connections ("visual chains") that establish correspondences between the same object parts across images. They show that the proposed localized approach better models relative attributes than baselines that either use global appearance features or stronger supervision.

Part III: Describing People Based on Attributes

Automatically describing people based on their fine-grained semantic attributes is important for many application domains, such as visual surveillance and e-commerce. The third part of the book covers state-of-the-art methods for estimation of human attributes and their use in different applications.

Chapter 8, by Chen Change Loy, Ping Luo, and Chen Huang, presents recent progress and cutting-edge methods based on deep learning for solving problems in estimating facial attributes such as gender, age, presence of facial hair, eyewear, hairstyle, and others. They cover approaches for handling class imbalance in attribute prediction, and demonstrate the use of facial attribute classification as an auxiliary task for improving face detection and face alignment.

In Chap. 9, Si Liu, Lisa Brown, Qiang Chen, Junshi Huang, Luoqi Liu, and Shuicheng Yan introduce methods that leverage facial and clothing attributes as a mid-level representation for applications related to fashion. In particular, they show that modeling attributes is crucial for fashion recommendation systems. In addition, they show that attributes play a major role in a system for clothing retrieval from online shopping catalogs.

Part IV: Defining a Vocabulary of Attributes

After covering multiple uses of visual attributes, as described earlier, we address the problem of discovering them, i.e., how to define a vocabulary of attributes.

In Chap. 10, Subhransu Maji surveys recent methods and defines a taxonomy of techniques for discovering a vocabulary of parts and attributes. The approaches discussed in this survey consider a vocabulary of attributes defined by experts and based on discovery methods, such as non-semantic embeddings, text mining, similarity comparisons, and others.

In Chap. 11, Genevieve Patterson and James Hays use crowdsourcing to generate a vocabulary of discriminative scene attributes related to affordances, materials, and spatial layout. After the attributes are discovered, they annotate more than ten thousand images with individual attribute labels, and show that attribute models derived from this data serve as an effective intermediate representation for zero-shot learning and image retrieval tasks.

Part V: Attributes and Language

We conclude our volume with a forward-looking topic: the connection of visual attributes and natural language.

In Chap. 12, Marcus Rohrbach discusses using visual attributes as semantic units between natural language and visual recognition. In particular, he covers methods for mining attributes from language resources, generating sentences from images and video, grounding natural language in visual content, and visual question answering.

In Chap. 13, Carina Silberer states that distributional models of word meaning have been criticized as "disembodied" in that they are not grounded in perception, and show that visual attributes predicted from images can be used as a way of physically grounding word meaning. Silberer introduces a new large-scale dataset of images annotated with visual attributes and a neural network-based model, which learns higher-level meaning representations by mapping words and images, represented by attributes, into a common embedding space.

References

1. Akata, Z., Malinowski, M., Fritz, M., Schiele, B.: Multi-cue zero-shot learning with strong supervision. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label embeddings for image classification. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **38**(7), 1425–1438 (2016)
3. Biederman, I.: Recognition by components—a theory of human image understanding. Psychol. Rev. **94**(2), 115–147 (1987)
4. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: European Conference on Computer Vision (ECCV) (2010)
5. Christie, G., Parkash, A., Krothapalli, U., Parikh, D.: Predicting user annoyance using visual attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
6. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)

8. Deza, A., Parikh, D.: Understanding image virality. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
9. Dhar, S., Ordonez, V., Berg, T.: High level describable attributes for predicting aesthetics and interestingness. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
10. Donahue, J., Grauman, K.: Image recognition with annotator rationales. In: International Conference on Computer Vision (ICCV) (2011)
11. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for efficient image retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
12. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
13. Feris, R.S., Bobbit, R., Brown, L., Pankanti, S.: Attribute-based people search: lessons learnt from a practical surveillance system. In: International Conference on Multimedia Retrieval (ICMR) (2014)
14. Gan, C., Yang, T., Gong, B.: Learning attributes equals multi-source domain generalization. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: International Conference on Computer Vision (ICCV) (2015)
16. Hendricks, L., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European Conference on Computer Vision (ECCV) (2016)
17. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: International Conference on Computer Vision (ICCV) (2015)
18. Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A.: What makes a photograph memorable? IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **36**(7), 1469–1482 (2014)
19. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: European Conference on Computer Vision (ECCV) (2016)
20. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: interactive image search with relative attribute feedback. Int. J. Comput. Vis. (IJCV) **115**, 185–210 (2015)
21. Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: International Conference on Computer Vision (ICCV) (2011)
22. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Conference on Neural Information Processing Systems (NIPS) (2012)
23. Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A search engine for large collections of images with faces. In: European Conference on Computer Vision (ECCV) (2008)
24. Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. In: ACM SIGGRAPH (2014)
25. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
26. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **36**(3), 453–465 (2013)
27. Layne, R., T., H., Gong, S.: Re-id: Hunting attributes in the wild. In: British Machine Vision Conference (BMVC) (2014)
28. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
29. Liu, L., Xu, H., Xing, J., Liu, S., Zhou, X., Yan, S.: Wow! you are so beautiful today! In: International Conference on Multimedia (ACM MM) (2013)
30. Palatucci, M., Hinton, G., Pomerleau, D., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Conference on Neural Information Processing Systems (NIPS) (2009)
31. Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision (ICCV) (2011)
32. Patterson, G., Hays, J.: Sun attribute database: discovering, annotating, and recognizing scene attributes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

33. Qiao, R., Liu, L., Shen, C., van den Hengel, A.: Less is more: zero-shot learning from online textual documents with noise suppression. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
34. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where and why? Semantic relatedness for knowledge transfer. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
35. Schein, A., Popescul, A., Ungar, L., Pennock, D.: Methods and metrics for cold-start recommendations. In: International Conference on Research and Development in Information Retrieval (ACM SIGIR) (2002)
36. Sharmanska, V., Quadrianto, N., Lampert, C.: Learning to rank using privileged information. In: International Conference on Computer Vision (ICCV) (2013)
37. Shi, Z., Hospedales, T., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
38. Siddiquie, B., Feris, R.S., Davis, L.: Image ranking and retrieval based on multi-attribute queries. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
39. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: European Conference on Computer Vision (ECCV) (2016)
40. Vaquero, D., Feris, R.S., Brown, L., Hampapur, A.: Attribute-based people search in surveillance environments. In: Winter Conference on Applications of Computer Vision (WACV) (2009)
41. Yan, X., Yang, J., Sohn, K., Le, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision (ECCV) (2016)
42. Yang, S., Luo, P., Loy, C., Tang, X.: From facial part responses to face detection: a deep learning approach. In: International Conference on Computer Vision (ICCV) (2015)
43. Zhang, Z., Luo, P., Loy, C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **38**(5), 918–930 (2015)

Part I
Attribute-Based Recognition

Chapter 2
An Embarrassingly Simple Approach to Zero-Shot Learning

Bernardino Romera-Paredes and Philip H. S. Torr

Abstract Zero-shot learning concerns learning how to recognise new classes from just a description of them. Many sophisticated approaches have been proposed to address the challenges this problem comprises. Here we describe a zero-shot learning approach that can be implemented in just one line of code, yet it is able to outperform state-of-the-art approaches on standard datasets. The approach is based on a more general framework which models the relationships between features, attributes, and classes as a network with two linear layers, where the weights of the top layer are not learned but are given by the environment. We further provide a learning bound on the generalisation error of this kind of approaches, by casting them as domain adaptation methods. In experiments carried out on three standard real datasets, we found that our approach is able to perform significantly better than the state of the art on all of them.

2.1 Introduction

Zero-shot learning (ZSL) is a relatively recent machine learning paradigm that was introduced in the works [21, 28], and quoting the latter, it aims to tackle the following question:

> Given a semantic encoding of a large set of concept classes, can we build a classifier to recognise classes that were omitted from the training set?

That is, ZSL consists in recognising new categories of instances without training examples, by providing a high-level description of the new categories that relate them to categories previously learned by the machine. This can be done by means of

B. Romera-Paredes (✉) · P.H.S. Torr
Department of Engineering Science, University of Oxford, Parks Road,
Oxford OX1 3PJ, UK
e-mail: bernard@robots.ox.ac.uk; bernardino.romeraparedes@eng.ox.ac.uk

P.H.S. Torr
e-mail: philip.torr@eng.ox.ac.uk

R.S. Feris et al. (eds.), *Visual Attributes*, Advances in Computer Vision
and Pattern Recognition, DOI 10.1007/978-3-319-50077-5_2

learning an intermediate encoding describing each class, referred to as attributes. In words of [1]:

> Attributes correspond to high-level properties of the objects which are shared across multiple classes, which can be detected by machines and which can be understood by humans.

One recurrent example that we mention in this chapter is the use of attributes such as *white*, *strong*, *furry*, and *quadrupedal*, to describe and learn classes of animals.

Zero-shot learning has attracted considerable attention due to both its wide applicability to many real- world situations and the singular challenges it presents. An example of ZSL happens when dealing with an ever growing set of classes, such as detecting new species of living beings, using attributes such as the ones mentioned in the previous example. Another scenario occurs when the granularity of the description of the categories to be distinguished makes it infeasible to obtain training instances for each of them, e.g. when a user wants to recognise a particular type of shoe (we refer to Chap. 9 for more on this topic). The main challenge ZSL poses is to design a model able to exploit the relations between features, attributes, and classes, so that the knowledge learned at the training stage can be transferred to the inference stage, in a similar way as human beings are able to understand a new concept, if it is described as a combination of previously known attributes or concepts [27]. Hereafter, we use the term signature to refer to this attribute description of a class.

Zero-shot learning is inherently a two-stage process: training and inference. In the training stage, knowledge about the attributes is captured, and in the inference stage this knowledge is used to categorise instances among a previously unseen set of classes. Many efforts have been made to improve the training stage [10, 15, 17], whereas the inference stage has received little attention [16]. For example many approaches blindly assume that all attributes convey the same amount of information, and can be predicted with the same accuracy, thus, they are evenly utilised in the inference rule. However these assumptions rarely hold true in real world cases.

We study a framework that is able to integrate both stages, overcoming the need to make strong and unrealistic assumptions, as the ones previously described. This framework, introduced in [1], is based on modelling the relationship between features, attributes, and classes as a (linear) model composed of two layers. The first layer contains the weights that describe the relationship between the features and the attributes, and is learned at the training stage. The second layer models the relationship between the attributes and the classes and is fixed using the prescribed attribute signatures of the classes. Given that the seen classes and the unseen classes are different, this second layer is interchangeable.

The main contributions of this work are:

- Given the framework in [1], we derive a principled choice of the regularizer, which has three nice properties:

  1. It performs comparably or better than the state of the art.
  2. It is efficient both at the training and at the inference stages.
  3. It is extremely easy to implement: one line of code for training and another one for inference (without calling any external functions).

- We provide a bound on the generalisation error of the approaches comprised in this framework. This is done by bridging the gap between zero-shot learning and domain adaptation, and making use of previous results in the latter [4, 5].

The remainder of the chapter is organised as follows. In Sect. 2.2 we briefly review methods proposed to deal with zero-shot learning. In Sect. 2.3 we describe the above ZSL framework, and present our method. In Sect. 2.4 we analyse its learning capabilities. In Sect. 2.5 we report the results of our experiments on one synthetic and three standard real datasets. Finally in Sect. 2.6 we discuss the main contributions of this work and propose several research lines that can be explored.

2.2 Related Work

Zero-shot learning relies on learning how to recognise several properties or attributes from objects, so that these learned attributes can be harnessed when used in the description of new, unseen classes. Indeed, it is attributes learning that drives the possibility of learning unseen classes based only on their description [27]. Within the context of machine learning, an antecedent of the notion of attribute learning can be found in [9] in the form of binary descriptors. The aim was using these binary descriptors as error-correcting codes, although these did not convey any semantic meaning. Recently, there has been an increasing interest in attributes learning, partially due to the availability of data containing tags or meta-information. This has proved to be particularly useful for images [10, 11, 21], as well as videos [13, 24].

Many papers focus on attributes learning, namely the training stage in zero-shot learning methods, putting special emphasis on the need to disentangle the correlations between attributes at the training stage, because these properties may not be present in the target data [17]. For example in [10] the authors focus on the feature extraction process with the aim of avoiding confusion in the learning process of attributes that often appear together in the training set instances.

With regard to the inference stage in which the predicted attributes are combined to infer a class, many approaches are variants of 1-nearest neighbour, or probabilistic frameworks. Approaches that resemble 1-nearest neighbour consist in looking in the attribute space for the closest unseen class signature to the predicted attribute signature of the input instance. It is used in [10], and in [28] the authors study risk bounds of this approach when using the Hamming distances between the predicted signature and the signatures of the unseen classes. Whereas 1-nearest neighbour is an intuitive way for inferring classes from the attributes, it presents several drawbacks. Namely, it treats equally all dimensions of the attribute space, which may be suboptimal, as some attributes are more important than others for discriminating between classes, and metrics such as Hamming distance ignore quantitative information in the prediction of the attributes.

In [21, 22] the authors propose a two-stage probabilistic framework in which the predictions obtained in the first stage can be combined to determine the most likely unseen class. Within this framework two approaches are proposed: directed attribute prediction (DAP), and indirect attribute prediction (IAP). In DAP a probabilistic classifier (e.g. logistic regression model) is learned at the training stage for each attribute. At the inference stage, the previous estimators are used to infer among the unseen classes provided their attributes signatures. In IAP one probabilistic classifier is learned for each seen class, whereas at the inference stage the predictions are combined accounting for the signatures of both seen and unseen classes. The DAP approach has been widely used by many other methods. In [35] the authors extend DAP by weighting the importance of each attribute, based on its frequency of appearance. These probabilistic approaches bring a principled way of combining the attribute predictions of a new instance in order to infer its class. However, in addition to being unable to estimate the reliability of the predicted attributes, they introduce a set of independence assumptions that hardly ever hold in real world, for example, when describing animals the attributes "terrestrial" and "farm" are dependent, but are treated as independent in these approaches.

Very recently, the authors of [16] proposed an approach that acknowledges uncertainty in the prediction of attributes, having mechanisms to deal with it. The approach is based on random forests that classify attribute signatures into the unseen classes, using a validation partition from the training set. The resultant model empirically proves to be superior to previous inference methods, such as DAP, and it obtains state-of-the-art results in the benchmark datasets. One of the limitations of this model is the need to have the attribute signatures of the unseen classes at the training stage. In other words, the model learned at the training stage is tailored to work with a predefined set of unseen classes.

The approach we describe in Sect. 2.3 bypasses the limitations of these methods by expressing a model based on an optimisation problem which relates features, attributes and classes. There are some works which follow a similar strategy. A relevant approach is the one described in [1], where the authors propose a model that implicitly learns the instances and the attributes embeddings onto a common space where the compatibility between any pair of them can be measured. The approach we describe here is based on the same principle, however we use a different loss function and regularizer which not only makes the whole process simpler and efficient, but also leads to much better results. Another related approach is proposed in [14], where the authors use the information regarding the correlations between attributes in both training and test instances. The main differences are that they focus on attribute prediction, and they employ a max-margin formulation that leads to a more complex approach. These approaches [1, 14], as well as the one we propose, can be seen as particular instances of the general framework described in [37], which unifies a wide range of multitask learning and multi-domain learning methods.

Other approaches consider the attributes as latent variables to be learned. For example in [36], an explicit feature map is designed to model the relationships

between features, attributes and classes. Other approaches, such as [24, 26], consider different schemes where attributes representations are to be learned.

The approach we describe is grounded on the machine learning areas of transfer learning and domain adaptation. The term transfer learning encompasses several machine learning problems, and has received several names, such as learning to learn [23] or inductive transfer [7, 31, 33]. Here, we refer to transfer learning in the lifelong learning sense, that is, the aim is to extract knowledge from a set of source tasks, so that it can be applied to learn future tasks more efficiently. Zero-shot learning problems share the necessity to extrapolate the knowledge gained previously to tackle a new learning scenario. The main difference is that in transfer learning the information about the new tasks is given as a set of labelled instances, whereas in zero-shot learning this information takes the form of descriptions of the unseen classes. An extensive review of transfer learning methods can be found in [29].

The aim of domain adaptation is to learn a function from data in one domain, so that it can be successfully applied to data from a different domain [4, 8, 19]. It resembles transfer learning but there are important differences to note. In transfer learning the marginal input distribution (domain) in both source and target tasks is supposed to be the same, whereas each task comprises a different objective predictive function. For example, given a set of journal documents sampled from a fixed marginal distribution, a source task may consist in classifying documents between different topics, and the target task could be about classifying each document in terms of its author. Domain adaptation makes the reverse assumption, that is, the objective predictive function is the same but the marginal distributions for source and target tasks are different. Following the previous example, now we have a common function to learn: classifying documents in terms of different topics. However the source and target tasks receive documents from two different journals, that is, from two different marginal distributions. The link between our approach and domain adaptation becomes clear in Sect. 2.4.1.

2.3 Embarrassingly Simple ZSL

In order to explain our approach, we start by describing a standard linear supervised learning method, and then extend that model to tackle the ZSL scenario. In the following, we adopt the convention of using lower-case letters to denote scalars, lower-cases bold letters to denote vectors, and higher-case bold letters to denote matrices.

Supervised linear model

Let us denote by $\mathbf{X} \in \mathbb{R}^{d \times m}$ the instances available at the training stage, where d is the dimensionality of the data, and m is the number of instances. Similarly we use

$\mathbf{Y} \in \{0, 1\}^{m \times z}$ to denote the ground truth labels of each training instance belonging to any of the z classes. In most cases, each row of $\mathbf{Y}$ contains only one positive entry indicating the class it belongs to. Nevertheless, the present framework allows an instance to belong to several classes simultaneously.

If we were interested in learning a linear predictor for z classes, we would optimise the following problem:

$$\underset{W \in \mathbb{R}^{d \times z}}{\text{minimise}} \; L\left(\mathbf{X}^\top \mathbf{W}, \mathbf{Y}\right) + \Omega\left(\mathbf{W}\right), \tag{2.1}$$

where $\mathbf{W}$ contains the parameters to be learned, L is a convex loss function, and Ω a convex regularizer. Problem (2.1) encompasses several approaches, depending on the choice of L and Ω. For example if L is the sum of hinge losses, and Ω is the Frobenius norm, this would lead to a standard support vector machine (SVM), but one can consider other loss functions such as logistic loss, and other regularizers, such as the trace norm, leading to multitask learning methods [2, 32].

ZSL model

Quoting [21], the formal definition of the ZSL problem can be described as follows:

> Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \subset \mathcal{X} \times \mathcal{Y}$ be training samples where $\mathcal{X}$ is an arbitrary feature space and $\mathcal{Y}$ consists of z discrete classes. The task is to learn a classifier $f : \mathcal{X} \longrightarrow \mathcal{Y}'$ for a label set $\mathcal{Y}'$ of z' classes, that is disjoint from $\mathcal{Y}$.

In order to accomplish that, we are given the attributes of all classes as additional information. We assume that each class is described by a known signature composed of a attributes. We can represent the training signatures in a matrix $\mathbf{S} \in [0, 1]^{a \times z}$. This matrix may contain boolean entries, when the description of classes is defined as a list of attributes, or more generally, it may contain for each attribute any value in $[0, 1]$ providing a soft link between attributes and classes. Together matrices $\mathbf{Y}$ and $\mathbf{S}$ provide enough information so that one can obtain the ground truth attributes for each instance.

In problem (2.1) the attributes are not used, and therefore, there is no way to perform knowledge transfer from this set of classes to new classes. One can introduce the given information about the attributes, $\mathbf{S}$, by introducing a mapping from the attributes to the feature space, $\mathbf{V}$, such that $\mathbf{W} = \mathbf{VS}$, where $\mathbf{V} \in \mathbb{R}^{d \times a}$. That leads to the following problem, similar to the one proposed in [1]:

$$\underset{\mathbf{V} \in \mathbb{R}^{d \times a}}{\text{minimise}} \; L\left(\mathbf{X}^\top \mathbf{VS}, \mathbf{Y}\right) + \Omega\left(\mathbf{VS}\right). \tag{2.2}$$

At the inference stage, given the features of an instance, $\mathbf{x} \in \mathbb{R}^d$, we wish to determine to which class it belongs to, among a new set of z' unseen classes, $\mathcal{Y}'$, disjoint from

the set of seen classes, $\mathcal{Y}$. To do so, we are provided with their attributes signatures, $\mathbf{S}' \in [0, 1]^{a \times z'}$. The prediction is then given by

$$\underset{i \in [1,\ldots,z']}{\operatorname{argmax}} \ \mathbf{x}^\top \mathbf{V} \mathbf{s}'_i, \tag{2.3}$$

where $\mathbf{s}'_i \in [0, 1]^a$ denotes the i-th column of matrix $\mathbf{S}'$.

One interpretation of this model is provided in [1]. There, each class is represented in the attribute space by means of its signature. Thus, the learning weights, $\mathbf{V}$, map any input instance, $\mathbf{x}$, into this attribute space. Given that both classes and instances are mapped into a common space, one can estimate the *compatibility* between them. Thus, at the inference stage, the model predicts the class in $\mathcal{Y}'$ that is most compatible with the input instance, by making use of (2.3). Note that if all given signatures are normalised, $\|\mathbf{s}'_1\|_2 = \|\mathbf{s}'_2\|_2 = \ldots \|\mathbf{s}'_{z'}\|_2$, then the notion of maximum compatibility among the signatures corresponds to finding the minimal Euclidean distance with respect to $\mathbf{V}^\top \mathbf{x}$ in the attribute space.

It is important to note the advantage of this model with respect to typical ZSL approaches reviewed in Sect. 2.2. Recall that these approaches were based on first estimating the attributes of a given instance, and then finding the class that best matches the predicted attributes, using some probabilistic or distance measure. In this way, all attributes are assumed to convey the same amount of information, an assumption that is likely detrimental, as often some attributes have more discriminative power than others. On the other hand, the approach in (2.2) is able to learn and exploit the relative importance of each of the attributes for discriminating between classes. For example, if the i-th attribute has less discriminative powers than the others, then the i-th column of the learned weights $\mathbf{V}$ should have a smaller norm than the others, so that it has a smaller contribution in the classification decision.

The method above makes the implicit assumption that for each attribute, its reliability to discriminate between seen classes is similar to its reliability to distinguish between unseen classes. In order to explain why this assumption is reasonable, let us recall the example of animals classification, and let us assume that we are given the attributes *it has teeth*, and *is white*. The former attribute may be more difficult to recognise than the latter, given that some instances of animals may not show the mouth, whereas the colour of an animal is easy to infer. Hence the importance of the attribute *it has teeth* for the final classification decision should be low, independently of the classes at hand, given that it is more difficult to learn a reliable predictor for that attribute. This assumption is relevant whenever the reliability on estimating the attributes remain constant, regardless of the classes considered. The key point of this framework is that it does not try to minimise explicitly the classification error of the attributes, which are an intermediate layer that we are not directly interested in. Instead, it minimises the multiclass error of the final classes, by both learning implicitly how to recognise attributes, and also pondering the importance of each of them in the decision of the class.

There are several points to note from problem (2.2). First, if the regularizer Ω is of the form $\Omega(\mathbf{B}) = \Psi\left(\mathbf{B}^\top \mathbf{B}\right)$ for an appropriate choice of the function Ψ, then by using the representer theorem [3], this leads to a kernel version of the problem, where only inner products between instances are used:

$$\underset{\mathbf{A}\in\mathbb{R}^{m\times a}}{\text{minimise}}\ L\left(\mathbf{KAS}, \mathbf{Y}\right) + \Psi\left(\mathbf{S}^\top \mathbf{A}^\top \mathbf{KAS}\right), \tag{2.4}$$

where $\mathbf{K} \in \mathbb{R}^{m\times m}$ is the Gram matrix, $k_{i,j} = \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \right\rangle$, being $\phi(\mathbf{x})$ the representation of $\mathbf{x}$ in a given feature space. Secondly, problem (2.2) and its equivalent problem (2.4) are convex, thus its globally optimal solution can be found.

A scheme of this framework is shown in Fig. 2.1. This framework is utilised in its linear form (Eq. 2.2) in [1], for a particular choice of the loss function (based on the hinge loss function), and the regularizer (based on the Frobenius norm of the learning weights). In the following, we describe and justify a different choice for those elements, which leads to a more efficient and effective training model.

2.3.1 Regularisation and Loss Function Choices

We now come to the first contribution of this chapter. The framework described above comprises several approaches, which vary depending on their regularizers and loss functions. Here we design a regularizer which accomplishes the following desiderata:

- Given any (training) attribute signature, $\mathbf{s}_i \in [0, 1]^a$ for some $i \in [1, \ldots, z]$, its mapping to the d-dimensional feature space is given by $\mathbf{V}\mathbf{s}_i \in \mathbb{R}^d$. This representation must be controlled so that ideally the mapping of all signatures on the feature space have a similar Euclidean norm. This allows fair comparisons between signatures, and prevents problems that stem from highly unbalanced training sets.
- Conversely, the mapping of each training instance $\mathbf{x}_i$, for $i \in [1, \ldots, m]$, into the a-dimensional attribute space is given by $\mathbf{V}^\top \mathbf{x}_i \in \mathbb{R}^a$. Similarly to the previous point, it would be interesting to bound the Euclidean norm of that term. The aim here is to map all instances to a common region in the attribute space. In this way, we can encourage the generalisation of the model to test instances, if their representation into the attribute space fall into the same region where the training instances lie.

A regularizer that accomplishes the previous points can be written as follows:

$$\Omega\left(\mathbf{V}; \mathbf{S}, \mathbf{X}\right) = \gamma \left\|\mathbf{VS}\right\|_{\text{Fro}}^2 + \lambda \left\|\mathbf{X}^\top \mathbf{V}\right\|_{\text{Fro}}^2 + \beta \left\|\mathbf{V}\right\|_{\text{Fro}}^2, \tag{2.5}$$

where the scalars γ, λ, β are the hyper-parameters of this regularizer, and $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm. The first two terms account for the above points, and we have added one further term consisting in a standard weight decay penalising the Frobenius norm of the matrix to be learned.

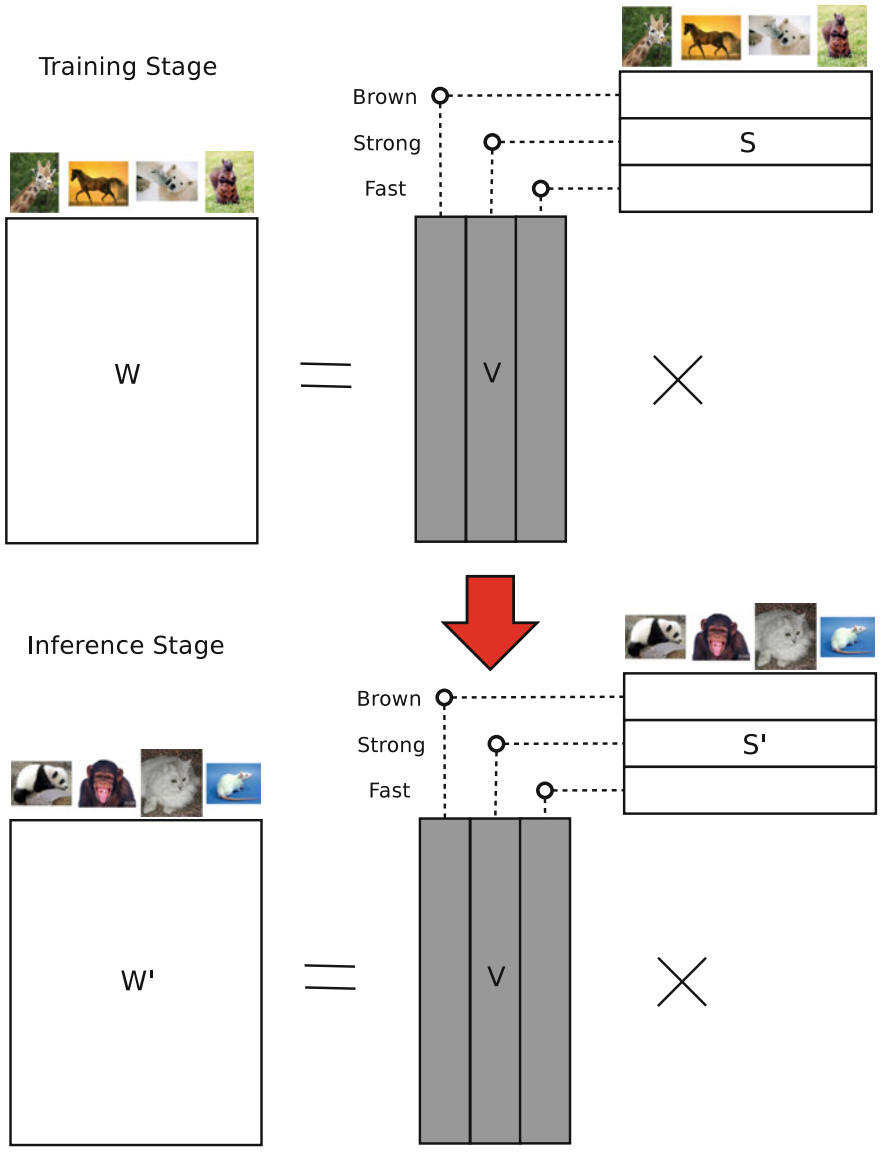

Fig. 2.1 Summary of the framework described in Sect. 2.3. At the training stage, we use the matrix of signatures $\mathbf{S}$ together with the training instances to learn the matrix $\mathbf{V}$ (in *grey*) which maps from the feature space to the attribute space. At the inference stage, we use that matrix $\mathbf{V}$, together with the signatures of the unseen classes, $\mathbf{S}'$, to obtain the final linear model $\mathbf{W}'$

Having made these choices, we note that if:

- $L(\mathbf{P}, \mathbf{Y}) = \|\mathbf{P} - \mathbf{Y}\|_{\text{Fro}}^2$.
- $\beta = \gamma\lambda$

then the solution to problem (2.2) can be expressed in closed form:

$$\mathbf{V} = \left(\mathbf{X}\mathbf{X}^\top + \gamma\mathbf{I}\right)^{-1} \mathbf{X}\mathbf{Y}\mathbf{S}^\top \left(\mathbf{S}\mathbf{S}^\top + \lambda\mathbf{I}\right)^{-1}. \tag{2.6}$$

This, and the corresponding kernel version that can be derived from (2.4), are the one-line-of-code solutions we mentioned in the introduction.

2.4 Risk Bounds

In this section we provide some theoretical guarantees about our approach, bounding the expected error on the inference stage with respect to the training error. In order to do so, we first transform our problem into a domain adaptation one.

2.4.1 Simple ZSL as a Domain Adaptation Problem

Let us assume that problem (2.2) can be expressed in the following way:

$$\underset{\mathbf{V}\in\mathbb{R}^{d\times a}}{\text{minimise}} \sum_{i=1}^{m}\sum_{t=1}^{z} \ell\left(\mathbf{x}_i^\top \mathbf{V}\mathbf{s}_t^\top, y_{t,i}\right) + \Omega\left(\mathbf{V}\right), \tag{2.7}$$

where $\ell(\cdot,\cdot) : \mathbb{R} \times \{-1, 1\} \longrightarrow [0, 1]$. That implies that one instance may be classified to belong to zero, one, or more than one classes. Such an assumption may be realistic in some cases, for example when there are some instances in the training set that do not belong to any seen class. Then, problem (2.7) can be expressed in a more conventional form:

$$\underset{\mathbf{v}\in\mathbb{R}^{da}}{\text{minimise}} \sum_{i=1}^{m}\sum_{t=1}^{T} \ell\left(\widetilde{\mathbf{x}}_{t,i}^\top \mathbf{v}, y_{t,i}\right) + \Omega\left(\mathbf{v}\right), \tag{2.8}$$

where

$$\widetilde{\mathbf{x}}_{t,i} = \text{vec}\left(\mathbf{x}_i \mathbf{s}_t^\top\right) \in \mathbb{R}^{da}. \tag{2.9}$$

Note that at the inference time, given a new instance, $\mathbf{x}$, the predicted confidence of it belonging to an unseen class t with attribute signature $\mathbf{s}_t$, is given by $\widetilde{\mathbf{x}}_t^\top \mathbf{v} = \mathbf{v}^\top \text{vec}\left(\mathbf{x}\mathbf{s}_t^\top\right)$. Therefore, even if the original test instances $\mathbf{x}$ were sampled from the same distribution as the training instances, the transformation of them

using attributes signatures makes the training and test instances come from different distributions. Note also that in the current settings, we are learning a unique common function across domains. As a consequence, we are facing a domain adaptation problem.

2.4.2 Risk Bounds for Domain Adaptation

Domain adaptation has been analysed from a theoretical viewpoint in several works [4, 5]. Here we apply these developments to our problem.

In a domain adaptation problem we assume that the training instances are sampled from a source distribution $\mathcal{D}$, and the test instances are sampled from a target distribution $\mathcal{D}'$. Following the definition of [4], a function h is said to be a predictor if it maps from the feature space to $\{0, 1\}$, and f is the ground truth labelling function for both domains, mapping from the feature space to $[0, 1]$. Then the expected error of h with respect to the source distribution is defined as:

$$\varepsilon(h) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[|f(\mathbf{x}) - h(\mathbf{x})|\right],$$

and the expected error of h with respect to the target distribution, $\varepsilon'(h)$, is defined accordingly.

Theorem 2 in [4] states that given a hypothesis space $\mathcal{H}$ of VC-dimension $\bar{d}$, and sets $\mathcal{U}, \mathcal{U}'$ of $\bar{m}$ instances sampled i.i.d. from $\mathcal{D}$ and $\mathcal{D}'$, respectively, then with probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$\varepsilon'(h) \leq \varepsilon(h) + 4\sqrt{\frac{2\bar{d}}{\bar{m}}\left(\log\frac{2\bar{m}}{\bar{d}} + \log\frac{4}{\delta}\right)} + \alpha + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{U}, \mathcal{U}'\right), \tag{2.10}$$

where

- α is an upper-bound of $\inf_{h\in\mathcal{H}}\left[\varepsilon(h) + \varepsilon'(h)\right]$. In particular if the ground truth function f is contained in $\mathcal{H}$, then $\alpha = 0$.
- $d_{\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right)$ is known as the $\mathcal{A}$-distance between distributions $\mathcal{D}$ and $\mathcal{D}'$ over the subsets defined in $\mathcal{H}$ [20]:

$$d_{\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right) = 2\sup_{h\in\mathcal{H}}\left|P_{\mathcal{D}}(h) - P_{\mathcal{D}'}(h)\right|,$$

  where $P_{\mathcal{D}}(h)$ denotes the probability of any event in h, under the distribution $\mathcal{D}$. This is equivalent to the expected maximal accuracy achieved by a hypothesis in $\mathcal{H}$ separating the instances generated by the two different distributions $\mathcal{D}$ and $\mathcal{D}'$. In a similar vein, $\hat{d}_{\mathcal{H}}\left(\mathcal{U}_S, \mathcal{U}_T\right)$ is defined as the empirical distance between the samples $\mathcal{U}$ and $\mathcal{U}'$.

- $\mathcal{H}\Delta\mathcal{H}$ is the symmetric difference hypothesis space of $\mathcal{H}$ and it is defined as: $\mathcal{H}\Delta\mathcal{H} = \left\{h(x) \oplus h'(x) : h, h' \in \mathcal{H}\right\}$, $\oplus$ being the XOR operator. That is, a hypothesis g is in $\mathcal{H}\Delta\mathcal{H}$, if for a couple of hypothesis h,h' in $\mathcal{H}$, $g(x)$ is positive if and only if $h(x) \neq h'(x)$ for all x.

In our case $\mathcal{H}$ is the hypothesis space composed of all linear classifiers, $\bar{m} = mz$, and $\bar{d} = da + 1$. Let us assume that both train and test instances are sampled from the same distribution, $\mathcal{C}$. When we do the transformation specified in Eq. (2.9) using $\mathbf{S}$ and $\mathbf{S}'$ for the training and test instances, we end up having two different distributions, $\mathcal{D}$, and $\mathcal{D}'$ and we are interested in quantifying the $\mathcal{A}$-distance between them over our symmetric difference hypothesis space, $d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right)$. The assumption about both train and test instances are sampled from the same distribution (before the transformation) may not hold true in many cases, however it can be a fair approximation in the standard case where the contribution of the differences of training and test distributions of the feature spaces is negligible in comparison to the differences between $\mathbf{S}$ and $\mathbf{S}'$ when quantifying the distance between distributions $\mathcal{D}$ and $\mathcal{D}'$.

We observe two extreme cases. The first one contemplates the trivial scenario where $\mathbf{S} = \mathbf{S}'$, so that both distributions are similar and thus the distance is 0. In that case, if $\alpha = 0$, the bound given in Eq. (2.10) becomes equivalent to the Vapnik–Chervonenkis bound on a standard classifier. The second case arises when each attribute signature of the seen classes is orthogonal to each attribute signature of the unseen classes, that is, for each $i \in \{1 \ldots z\}$, $j \in \left\{1 \ldots z'\right\}$, $\left\langle \mathbf{s}_i, \mathbf{s}'_j \right\rangle = 0$.

To make the explanation of the latter case clearer let us denote by $\mathbf{x} \in \mathbb{R}^d$ any training instance in the original feature space, and similarly let $\mathbf{x}' \in \mathbb{R}^d$ be any test instance. Then, by applying equation (2.9) using the training signature $\mathbf{s}_i$, and test signature $\mathbf{s}'_j$ we have

$\widetilde{\mathbf{x}}_i = \text{vec}\left(\mathbf{x}\mathbf{s}_i^\top\right) \in \mathbb{R}^{da}$

$\widetilde{\mathbf{x}}'_j = \text{vec}\left(\mathbf{x}'\mathbf{s}_j'^\top\right) \in \mathbb{R}^{da}$

Note that because of the orthogonality assumption between training and test signatures the following holds true:

$$\left\langle \widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}'_j \right\rangle = \text{trace}\left(\mathbf{x}\mathbf{s}_i^\top \mathbf{s}'_j \mathbf{x}'^\top\right) = 0. \tag{2.11}$$

Equation (2.11) implies that in the new feature space any training instance is orthogonal to any test instance. Because of that, the following lemma becomes useful.

Lemma 1 *Let us consider $\mathcal{H}$ be the hypothesis space composed of all linear classifiers. Then given two orthogonal sets $\mathcal{P}$, $\mathcal{Q}$, in which the element 0 is not in either of them, there exists a hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ which separates them.*

Proof Let us consider any couple of points $\mathbf{p} \in \mathcal{P}$, $\mathbf{q} \in \mathcal{Q}$ with the only condition that they are not zero. We define

$h(\mathbf{x}) = \text{sign}\left((\mathbf{p} + \mathbf{q})^\top \mathbf{x}\right)$, and

$h'(\mathbf{x}) = \text{sign}\left((\mathbf{p} - \mathbf{q})^\top \mathbf{x}\right)$.

For any point $\mathbf{p}' \in \mathcal{P}$, $h(\mathbf{p}') = h'(\mathbf{p}')$, given that by definition $\mathbf{p}'$ and $\mathbf{q}$ are orthogonal. Similarly, for any point $\mathbf{q}' \in \mathcal{Q}$, $h(\mathbf{q}') = -h'(\mathbf{q}')$.

Therefore, for any point in $\mathcal{Q}$, $g \in \mathcal{H}\Delta\mathcal{H}$ associated to functions $h, h' \in \mathcal{H}$ will be positive, and for any point in $\mathcal{P}$, the same function g will be negative. □

As a consequence of Lemma 1, when the orthogonality assumption holds, the right-hand side term in Eq. (2.10) becomes bigger than 1, so that the bound is vacuous. One illustrative instance of this case happens when $\mathbf{S} = [\mathbf{B}, \mathbf{0}^{\mathbf{a},\mathbf{c}},]$, and $\mathbf{S}' = \left[\mathbf{0}^{\mathbf{a},\mathbf{b}}, \mathbf{C}\right]$ for some non-zero matrices $\mathbf{B} \in \mathbb{R}^{a\times b}$, $\mathbf{C} \in \mathbb{R}^{a\times c}$. In that case, the set of attributes that describe the seen classes are completely different from the ones describing the unseen classes, thus no transfer can be done.

All real scenarios lay between the previous cases. One interesting question is to characterise the value $d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right)$ as a function of solely $\mathbf{S}$ and $\mathbf{S}'$. We leave this question open.

2.5 Experiments

In order to assess our approach and the validity of the statements we made, we conducted a set of experiments on one synthetic and three real datasets, which comprise a standard benchmark of evaluation of zero-shot learning methods.[1]

2.5.1 Synthetic Experiments

First we used synthetically generated data with the aim of both checking the correctness of the described method, which we refer to as ESZSL (embarrassingly simple zero-shot learning), and comparing it with the baseline algorithm DAP on a controlled set-up. All hyper-parameters required by these methods were tuned by a validation process. This process is based on leaving out one subset of validation classes, so that the performance of the model is validated against them. In all cases the range of values tried for the hyper-parameters was 10^b, for $b = -6, -5, \ldots, 5, 6$. This set of values was chosen after performing preliminary experiments which empirically showed that the optimal performance for both approaches is found within this interval.

The data were generated as follows. Initially, we created the signatures for the classes by sampling each element of $\mathbf{S}$ from a Bernoulli distribution with 0.5 mean. We created the ground truth mapping from the attributes to the features, $\mathbf{V}^+ \in \mathbb{R}^{a\times d}$, where we have fixed $a = 100$ and $d = 10$, by sampling every element of it from a Gaussian distribution $\mathcal{G}(0, 1)$. The value of d is intentionally low so that there appear correlations between the attributes, as is usually the case in real data. For each class t,

[1] The code can be found at http://romera-paredes.com/zsl.

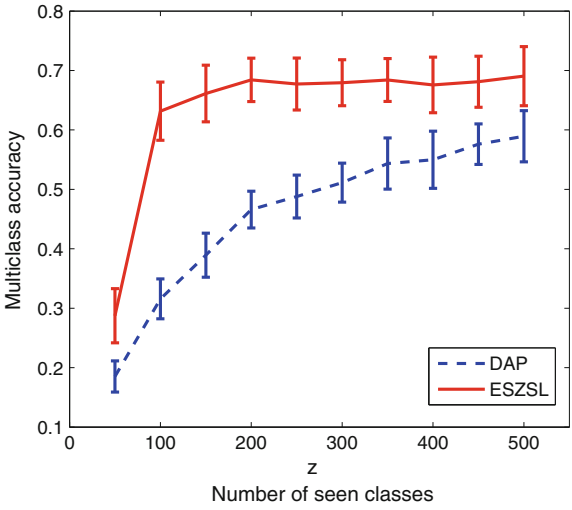

Fig. 2.2 Multiclass accuracy obtained by DAP [21], and ESZSL (Sect. 2.3.1), when varying the number of seen classes, z. *Vertical bars* indicate ± 1 standard deviation

we created 50 instances by first generating their representation in the attribute space by adding Gaussian noise, $\mathcal{G}(0, 0.1)$ to the attribute signature $\mathbf{S}_t$, then we brought them back onto the original feature space by using $\mathbf{V}^+$. Following this process, we generated a training set composed of z seen classes, and a test and validation set composed of 100 unseen classes each.

In the first experiment, we evaluated how the number of seen classes affected the performance of the methods on unseen classes. To do so, we varied the number of seen classes from 50 to 500 in intervals of 50. According to the results shown in Fig. 2.2, we can see that ESZSL significantly outperforms DAP in all cases. It is remarkable that the performance of ESZSL with 100 seen classes is superior to the performance of DAP with 500 seen classes. We also observe that the performance of ESZSL plateaus when the number of seen classes is above 200, possibly because there is no further margin of improvement.

In Sect. 2.3 we argue that the described approach should be robust to attributes having different discriminative capabilities for characterising the classes. In the second experiment, we assess how the approaches perform in the extreme case where some attributes provide no information at all about the classes at hand. The way we have implemented this is by first, synthesising a dataset just as described above, and second, by randomly selecting a set of attributes (without replacement) so that their information in all signatures is corrupted. In particular let us define by $\mathcal{A}$ the set of all attributes, with cardinality $|\mathcal{A}| = a$. From this set $\mathcal{A}$ we randomly sample ψ misleading attributes, creating the set $\Psi \subseteq \mathcal{A}$, $|\Psi| = \psi$. The way each of the inputs of the attributes in Ψ is corrupted is again by sampling from a Bernoulli distribution with 0.5 mean. In this experiment we have tried different values of ψ in the range of 5–45 attributes (out of 100), in intervals of 5. The results, reported in Fig. 2.3, show that our method significantly outperforms the baseline. For example we observe that when