To Know and To Learn
About the Integration of Knowledge Representation and Deep Learning for Fine-Grained Visual Categorization

Francesco Setti
Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy

Keywords: Fine-Grained Visual Categorization, Knowledge Representation, Deep Learning, Convolutional Neural Networks, Ontology.

Abstract: Fine-grained visual categorization is becoming a very popular topic for computer vision community in the last few years. While deep convolutional neural networks have been proved to be extremely effective in object classification and recognition, even when the number of classes becomes very large, they are not as good in handling fine-grained classes, and in particular in extracting subtle differences between subclasses of a common parent class. One way to boost performances in this task is to embed external prior knowledge into standard machine learning approaches. In this paper we will review the state of the art in knowledge representation applied to fine-grained object recognition, focusing on methods that use (or can potentially use) convolutional neural networks. We will show that many research works have been published in the last years, but most of them make use of knowledge representation in a very naïve (or even unaware) way.

1 INTRODUCTION

When we train a classifier, we want a machine to perform in few minutes, hours or days a learning process that for a human being takes years or decades. While the huge gain in speed is motivated by the computational power of a modern days computer, the assumption that we are mimicking the learning process of a human is mostly wrong. Indeed, the vast majority of classification methods are driven only by data, which is not the case of humans.

Our brain is able to generalize much better than machines, that easily fall into overfitting problems, and we are also able to identify inter-class connections that are usually neglected in automatic processes. More important, humans can learn to distinguish new visual classes by means of non-visual knowledge, like semantic information. Think, for example, to a child that sees for the first time a duck, he is not able to say it’s a duck, but still he can clearly say it’s “a bird with a yellow flat beak”. Going back to his mother, he asks her which kind of bird it was, only saying it had a weird yellow flat beak; the mum’s answer is “If the body was white, it’s a goose; but if it was grey with a green head, it is definitely a duck!”.

From now on, the child is not only able to recognise a duck, but he is also able to identify a goose without having ever seen one.

To extend the standard learning procedure, and allow it to handle this kind of situations, many different approaches such as transfer learning (Ding and Fu, 2017), domain adaptation (Bergamo and Torresani, 2010), and zero-shot learning (Lampert et al., 2014) have been proposed. Eventually, all of them implicitly embed some sort of prior knowledge into the classifier, sometimes at a very naïve level (e.g. the number of classes), while in other cases more complex knowledge (like attributes or class taxonomy) are also considered.

Standard image recognition problems are performed at a coarse-level, aiming to distinguish between completely unrelated classes (e.g. birds, cars, and chairs). Fine-Grained Visual Categorization (FGVC) addresses the problem of recognising domain specific classes (e.g. different species of birds), where visual similarity becomes extremely high, and only few details allow to discriminate between two different categories. In such a task, objects belonging to different classes may have marginal appearance differences, while objects within the same category may present larger appearance variations due to changes of scales or viewpoints, complex backgrounds and occlusions (see Fig. 1). High inter-class variance, in conjunction with low intra-class variance poses a very tough chal-
Figure 1: Two species of gulls from CUB 200 dataset illustrate the difficulty of fine-grained object classification: large intra-class and small inter-class variance. (originally appeared on (Zhao et al., 2017)).

The goal of this paper is to analyse which kind of prior knowledge is used in the state-of-the-art approaches and how it is included inside the classification procedure. We will show that many different sources of information have been used in the last years, often in an implicit (or even unaware) way, and that formal knowledge representation methods are rarely employed.

2 WHICH INFORMATION CAN BE USED?

Several different types of information can be used to build a knowledge base. In particular, when structured information is already available in the literature of relevant domains (as in the case of biological data), it is possible to use formal logics to define a set of relations between classes and different entities. For instance, the easiest relation that is usually included in all the ontologies is the “IS A” relation, which allow to specify a hierarchical structure (i.e. a taxo-
nomony) of classes. Other common relations are the “IS PART OF”, that is tightly related to all the part based methods for image classification, and the “HAS COLOR”, “HAS TEXTURE”, and “HAS DIMENSION” relations that are linked with attributes reasoning. In the following we will present the most common additional information that are used in image classification.

Hierarchy of Classes

A hierarchical taxonomy is a tree structure of classes for a given set of objects; it can be formalized as a directed graph where nodes represent classes and edges represent parent-child relationships. This representation is at the base of the popular computational lexicon WordNet (Miller, 1995), and thus of the very popular large scale visual classification dataset ImageNet (Deng et al., 2009).

Hierarchical models explicitly represent category hierarchies that admit sharing the appropriate abstract knowledge across them via a prior abstracted from related classes. The basic idea is that a hawk is visually more similar to instances of the class bird than any other class like car or dog. Many different approaches have been proposed, spanning from compound architectures (Salakhutdinov et al., 2013) to multi-task learning (Fan et al., 2017) to cascade classifiers (Wang et al., 2015b; Wang et al., 2015a).

The knowledge of class hierarchies is often used in combination with attributes (Akata et al., 2016; Al-Halah and Stiefelhagen, 2015; Romera-Paredes and Torr, 2015). Alternatively the graph can be used to combine the response of multiple one-vs-all classifiers in terms of conditional probability (Deng et al., 2014; Rohrbach et al., 2011), or to combine different network responses in an ensemble of networks approach (Wang et al., 2015b; Wang et al., 2015a).

Attributes

Visual attributes are mid-level semantic properties that are shared across different categories. They are at the same time semantic (human-understandable) and visual (machine-detectable). This allows to instruct the classifier to consider specific features that are known to be discriminative between two classes, and to recognize novel unseen categories by leveraging visual attributes classifiers, learned on known categories, and a semantic description of the new class. While early works required a precise manual annotations of attribute-class relations (Lampert et al., 2014; Farhadi et al., 2009), recent studies rely on information mined from textual sources, often in connection with semantic embeddings such as distributional word vector representations (Akata et al., 2016; Demirel et al., 2017; Gan et al., 2016; Qiao et al., 2016).

Attributes have been proved to be extremely powerful in large-scale coarse-level image classification (Ouyang et al., 2015), but can also be beneficial in case of fine-grained object recognition (Branson et al., 2010; Yu and Grauman, 2017). The main problem with fine-grained classes is that discriminative features are usually very few and very localized, making the attribute detection extremely challenging. On the other hand, in this scenario structured prior knowledge is usually available from specific studies (e.g. accurate description of distinctive features of each bird species).

As a matter of method, attributes can be used as privileged information during training time (Sharmanska et al., 2013), in an active (Kovashka et al., 2011) or multitask (Chen et al., 2017) learning, or with humans in the loop (Kovashka et al., 2015).

Parts

The key advantages of exploiting part representations for object classification is that parts have lower intra-class variability than whole objects, they deal better with pose variation and their configuration provides useful information about the aspect of the object.

In general, semantic part-based models treat an object as a collection of parts that models its shape and appearance. This facilitates fine-grained categorization by explicitly isolating subtle appearance differences associated with specific object parts. Traditional works follow a framework based on three steps: first, parts are detected and localized inside the image, then, parts are aligned to generate a pose-normalized representation of the object, and lastly a classifier is applied to perform categorization.

Part-based methods can be grouped into two sets, considering whether or not a part has a semantic meaning. In the first case, a part is any patch that is discriminative for the object class recognition (Endres et al., 2013; Felzenszwalb et al., 2010; Juneja et al., 2013), typically discovered from training images automatically, without human supervision. In the latter, parts are semantic concepts (e.g. ‘head’, ‘beak’, ‘wing’) that are easily interpretable by humans (Branson et al., 2014; Lin et al., 2015; Zhang et al., 2013; Zhang et al., 2014). While the first approach is unable to handle new classes (zero-shot learning) and generates models that have no meaning for humans, existing works on semantic part detection require part location annotations in the training images, which are very expensive to obtain. To overcome this problem, (Mo-
dolo and Ferrari, 2017) propose to learn part models from images collected by web search engines; this work requires only a list of parts forming the object, then it collects images from the web and employs an incremental learning strategy to gradually move from parts to whole objects, learning in the meanwhile the spatial arrangement of the components.

Part knowledge is usually exploited in the design of the network architecture, applying parallel networks to identify different body parts that are finally composed in a last bank of fully connected layers (Branson et al., 2014; Huang et al., 2016).

3 HOW CAN WE USE THIS KNOWLEDGE?

Prior knowledge can be used to learn a classifier either at a data level, at an architectural level, or at a procedural level. In the first case, semantics and structured knowledge about the domain can be used to automatically generate training data. In the second case, the provided information are directly included into the deep network in a sequential (or hierarchical) or in a multi-task approach. In the last case, prior knowledge is used as privileged information for learning a model that won’t use it in the testing phase.

Training Data Generation

A good training set has to be sufficiently informative to capture the nature of the object under analysis, but at the same time has to be generic enough to avoid overfitting and to cope with new instances of the same class. The task of generating these data is very time consuming and error prone. To overcome these problems unsupervised and weakly supervised methods have been proposed, while on the other hand researchers also tried to automatize this annotation process. In (Setti et al., 2013) WordNet is used as knowledge base to identify a set of words semantically related with the target class label, these are then filtered to remove all the words unrelated to visual specifications, and lastly used to query a web image search engine. This work has been extended in (Cheng et al., 2015), where Google N-grams is used to retrieve also statistics about the frequency of appearance of related words in a text corpus, and improve the semantic filter. (Movshovitz-Attias et al., 2015) exploits an ontology of geographical concepts to automatically propagate business category information and create a large, multi-label, training dataset for fine grained storefront classification. A classifier based on GoogLeNet/Inception Deep Convolutional Network architecture is then trained on 208 categories, achieving human level accuracy.

Network Architecture

The key idea of (Rohrbach et al., 2011) is to take advantage of the structured nature of the object category space to transfer knowledge between classes; specifically, they uses the structure of WordNet to build a class taxonomy, then classifiers are trained for both leaves and inner nodes, and finally knowledge is propagated throughout the graph in terms of conditional probability. Despite this work is experimented by using Bag-of-Words features, there is no theoretical obstacle to using CNN features. In fact, (Al-Halah and Stiefelhagen, 2015) uses a similar approach with CNN features; they also transfer attribute labels across classes, learn the attributes that are most discriminative between similar categories, and select the best attributes to share with a novel class.

(Deng et al., 2014) introduces Hierarchy and Exclusion (HEX) graphs as a standalone layer that can be used on top of any feedforward architecture for classification. Differently from the standard taxonomies mentioned so far, HEX can capture three semantic relations between two labels applied to the same object: mutual exclusion, overlap and subsumption. In (Wang et al., 2015a) an ensemble of networks approach is used to build a cascade classifier able to recognize classes at different levels of granularity (e.g. bird, woodpecker, and acorn woodpecker), each one focusing on different salient regions that are learned during the training phase.

(Akata et al., 2016) introduced a framework based on modelling the relationship between features, attributes, and classes as a model composed of two layers: one mapping images (i.e. features) to attributes, and the second mapping attributes to class labels. (Romera-Paredes and Torr, 2015) extends the approach by forcing the weights of the second layer to be formalized by an external knowledge base, instead of learned from data.

Privileged Information

Privileged information is a type of information that is only available during training. A number of information can be treated as privileged such as text, attributes, and bounding boxes of object parts.

Textual information is used by (Sharmanska et al., 2013), where the meaningful data are manually defined by the designer according to the specific problem; this operation is time consuming and the knowledge acquired is hardly transferable to new problems. On
the contrary, (Chen and Zhang, 2017) learns a label embedding model to define the visual-semantic misalignment between image features extracted by a CNN and class labels; this is then used to compute the cost function to learn a SVM classifier. This system is completely automatic and can be easily extended to new classes by simply adding data, but it only uses the CNN as a feature extractor, without any interaction between privileged information and the network itself.

Indeed, this pipeline of using CNN for features extraction and SVM for classification is very common for learning using privileged information. This is mostly due to the fact that privileged information in SVM has been widely studied and SVM+ (Vapnik and Vashist, 2009) is a well established learning paradigm. This pipeline applies not only to text data, but also to attributes. (Sharmanska and Quadrianto, 2017) proposes two different methods to use semantic attributes as privileged information, showing that, even in the era of deep convolutional neural networks, semantic attributes are useful when dealing with challenging computer vision tasks.

4 CONCLUSIONS

In this paper we reviewed the state-of-the-art in knowledge representation applied to fine-grained object recognition, with a particular attention to those methods that use convolutional neural networks for classification or feature extraction. Despite many research works have been published in the last years, only few of them are actually aware of the power that formal knowledge representation has in this kind of task, while most of the researchers use these prior knowledge in a very naive way. We believe that more attention should be payed to this topic, also taking inspiration from current research in different fields like formal ontologies and natural language processing.

ACKNOWLEDGEMENTS

This work is funded by the EDGE Company\(^1\) within the Bird Concentration Monitoring System project.

REFERENCES


\(^1\)theedgecompany.net


Duan, K., Parikh, D., Crandall, D., and Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In CVPR. IEEE.


Huang, S., Xu, Z., Tao, D., and Zhang, Y. (2016). Part-stacked CNN for fine-grained visual categorization. In CVPR.


Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., and Zhang, Z. (2015a). Multiple granularity descriptors for fine-grained categorization. In ICCV.


