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

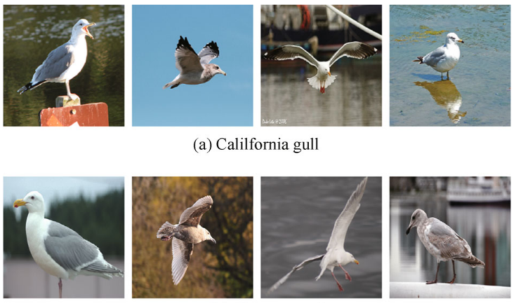
When we train a classifier, we want a machine to perform in few minutes, hours or days a learning process that for a human being takes years or decades. While the huge gain in speed is motivated by the computational power of a modern days computer, the assumption that we are mimicking the learning process 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-

Setti, F.
To Know and To Learn - About the Integration of Knowledge Representation and Deep Learning for Fine-Grained Visual Categorization.
DOI: 10.5220/0006651803870392
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 387-392
ISBN: 978-989-758-290-5



(a) Calilfornia gull

(b) Glaucous gull

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)).

lenge that is intrinsic for any FGVC problems. Moreover, in coarse-level object categorization, the background of an image often contributes significantly to the classication. This is not the case of FGVC, where in most of the cases background is shared across all the classes (e.g. background of flying birds is always sky, no matter what is the species of the bird); as a consequence, background is not that informative and it is usually a source of noise. FGVC studies have been conducted on motorcycles (Hillel and Weinshall, 2007), aircrafts (Maji et al., 2013), flowers (Rejeb Sfar et al., 2013), trees (Kumar et al., 2012), dogs (Khosla et al., 2011), butterflies (Duan et al., 2012), and birds (Berg et al., 2014).

Indeed, the biological domains are extremely well-suited to the problem. The reason is that centuries of biological studies lead to a biological classification (i.e. the Linnaean taxonomy) that dictates a clear set of mutually exclusive subcategories.

Deep learning techniques, and in particular deep neural networks, have received more and more attention for visual classification since 2012, when AlexNet (Krizhevsky et al., 2012) won the ILSVRC competition. Convolutional Neural Networks (CNN) are by far the most common type of deep networks in computer vision. CNN is a type of feed-forward artificial neural network in which the first layers are convolutional, i.e. the weights are entries of filters (or kernels) used to perform bi-dimensional convolution and each neuron processes data only for its receptive field. There are a huge number of variants of CNN architectures, but most of them are variants of 4 seminal works: AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), Inception modules (Szegedy et al., 2015) and ResNet (He et al., 2016). The core of the deep learning technology is that features are not designed by human engineers, but instead learned from data using a general-purpose learning procedure. This means the information is retrieved from data in a purely automatic way, with limited possibility for the designer to encode his knowledge into the machine.

As noted by (Berg et al., 2014), prior knowledge can dramatically improve the performance of classification systems, in particular for fine-grained classification. As an example, a spatio-temporal prior is attractive for bird species identification, because the density of bird species varies considerably across the world and throughout the year, due to migration; based on this observation, a system that can integrate images with information about location and date of acquisition would lead to better classification results. Unfortunately, in most of the public benchmarks only images and labels annotation (with or without bounding box location) are provided. Nevertheless, many research works managed to embed various types of prior knowledge into visual recognition methods. Some foundational works analyse the potential of using ontologies in computer vision (Porello et al., 2013; Town, 2006), and in the recent years these technologies have been applied to internet image search (Cheng et al., 2015; Setti et al., 2013), video-surveillance (Xu et al., 2015), video indexing (Benmokhtar and Huet, 2014), and remote sensing (Andrés et al., 2017). Surprisingly, despite ontologies are widely used for the generation of datasets (e.g. ImageNet is heavily based on WordNet, a semantic ontology), very few works explicitly adopt this formalism, while in many papers the ontological rules are informally used.

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 taxonomy) 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 ensamble 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 recognise 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 benefial 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 intraclass 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 wether 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, (Modolo 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 arrangment 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 leafs 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 leraning 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 naïve 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

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016). Label-embedding for image classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.

Al-Halah, Z. and Stiefelhagen, R. (2015). How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WACV*. IEEE.

Andrés, S., Arvor, D., Mougenot, I., Libourel, T., and Durieux, L. (2017). Ontology-based classification of remote sensing images using spectral rules. *Computers & Geosciences*, 102:158–166.

Benmokhtar, R. and Huet, B. (2014). An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, 73(2):663–689.

Berg, T., Liu, J., Woo Lee, S., Alexander, M. L., Jacobs, D. W., and Belhumeur, P. N. (2014). Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*.

Bergamo, A. and Torresani, L. (2010). Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*.

Branson, S., Van Horn, G., Belongie, S., and Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets. In *BMVC*.

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., and Belongie, S. (2010). Visual recognition with humans in the loop.

Chen, C.-Y., Jayaraman, D., Sha, F., and Grauman, K. (2017). Divide, share, and conquer: Multi-task attribute learning with selective sharing. In *Visual Attributes*, pages 49–85. Springer.

Chen, K. and Zhang, Z. (2017). Learning to classify fine-grained categories with privileged visual-semantic misalignment. *IEEE Trans. on Big Data*, 3(1):37–43.

Cheng, D. S., Setti, F., Zeni, N., Ferrario, R., and Cristani, M. (2015). Semantically-driven automatic creation of training sets for object recognition. *Computer Vision and Image Understanding*, 131:56–71.

Demirel, B., Cinbis, R. G., and Ikizler-Cinbis, N. (2017). Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *ICCV*.

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. (2014). Large-scale object classification using label relation graphs. In *ECCV*. Springer.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Ding, Z. and Fu, Y. (2017). Robust transfer metric learning for image classification. *IEEE Trans. on Image Processing*, 26(2):660–670.

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

Endres, I., Shih, K. J., Jiaa, J., and Hoiem, D. (2013). Learning collections of part models for object recognition. In *CVPR*.

Fan, J., Zhao, T., Kuang, Z., Zheng, Y., Zhang, J., Yu, J., and Peng, J. (2017). Hd-mtl: Hierarchical deep multi-task learning for large-scale visual recognition. *IEEE Trans. on Image Processing*, 26(4):1923–1938.

[1] theedgecompany.net

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *CVPR*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Gan, C., Yang, T., and Gong, B. (2016). Learning attributes equals multi-source domain generalization. In *CVPR*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hillel, A. B. and Weinshall, D. (2007). Subordinate class recognition using relational object models. In *NIPS*.

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

Juneja, M., Vedaldi, A., Jawahar, C. V., and Zisserman, A. (2013). Blocks that shout: Distinctive parts for scene classification. In *CVPR*.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshops*.

Kovashka, A., Parikh, D., and Grauman, K. (2015). Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210.

Kovashka, A., Vijayanarasimhan, S., and Grauman, K. (2011). Actively selecting annotations among objects and attributes. In *ICCV*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*. Springer.

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

Lin, D., Shen, X. Y., Lu, C. W., and Jia, J. Y. (2015). Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. (2013). Fine-grained visual classification of aircraft. Technical report.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of ACM*, 38(11):39–41.

Modolo, D. and Ferrari, V. (2017). Learning semantic part-based models from google images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PP(99):1–1.

Movshovitz-Attias, Y., Yu, Q., Stumpe, M. C., Shet, V., Arnoud, S., and Yatziv, L. (2015). Ontological supervision for fine grained classification of street view storefronts. In *CVPR*.

Ouyang, W., Li, H., Zeng, X., and Wang, X. (2015). Learning deep representation with large-scale attributes. In *ICCV*.

Porello, D., Setti, F., Ferrario, R., and Cristani, M. (2013). Multiagent socio-technical systems. an ontological approach. In *COIN Workshop, in conj. with AAMAS*.

Qiao, R., Liu, L., Shen, C., and van den Hengel, A. (2016). Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*.

Rejeb Sfar, A., Boujemaa, N., and Geman, D. (2013). Vantage feature frames for fine-grained categorization. In *CVPR*.

Rohrbach, M., Stark, M., and Schiele, B. (2011). Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*. IEEE.

Romera-Paredes, B. and Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *ICML*.

Salakhutdinov, R., Tenenbaum, J. B., and Torralba, A. (2013). Learning with hierarchical-deep models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(8):1958–1971.

Setti, F., Cheng, D.-S., Abdulhak, S. A., Ferrario, R., and Cristani, M. (2013). Ontology-assisted object detection: Towards the automatic learning with internet. In *ICIAP*. Springer.

Sharmanska, V. and Quadrianto, N. (2017). In the era of deep convolutional features: Are attributes still useful privileged data? In *Visual Attributes*, pages 31–48. Springer.

Sharmanska, V., Quadrianto, N., and Lampert, C. H. (2013). Learning to rank using privileged information. In *ICCV*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ICLR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR*.

Town, C. (2006). Ontological inference for image and video analysis. *Machine Vision and Applications*, 17(2):94–115.

Vapnik, V. and Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5):544–557.

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

Wang, Z., Wang, X., and Wang, G. (2015b). Learning fine-grained features via a cnn tree for large-scale classification. *arXiv preprint arXiv:1511.04534*.

Xu, Z., Liu, Y., Mei, L., Hu, C., and Chen, L. (2015). Semantic based representing and organizing surveillance big data using video structural description technology. *Journal of Systems and Software*, 102:217–225.

Yu, A. and Grauman, K. (2017). Fine-grained comparisons with attributes. In *Visual Attributes*, pages 119–154. Springer.

Zhang, N., Donahue, J., Girshick, R., and Darrell, T. (2014). Part-based R-CNNs for fine-grained category detection. In *ECCV*.

Zhang, N., Farrell, R., Iandola, F., and Darrell, T. (2013). Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*.

Zhao, B., Feng, J., Wu, X., and Yan, S. (2017). A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, pages 1–17.