A Review on Deep Learning Based Object Detection and Part Localization

[1]R.Fathima Syreen,[2]Dr.K.Merriliance

1.Research Scholar,Assistant Professor,Dept of Computer Application,Sadakathullah Appa College,Tirunelveli

2.Assistant professor,Dept of MCA,Sarah Tucker College,Tirunelveli.

Abstract—**The rapid growth of image data leads to the need of research and development of image retrieval. Compared to normal image, fine grained images are difficult to classify. Recognizing objects in fine grained domains can be extremely challenging task due to subtle differences and variations in poses, scales and rotations between same species. Deep Learning has recently achieved superior performance on many tasks such as image classification, object detection and neural language processing. In this survey we mainly focus on the object detection, semantic part localization and feature extraction which can facilitate fine grained categorization.**

Keywords—fine grained images;Deep Learning; object detection;part localization.

I. Introduction

Fine Grained object classification aims to distinguish objects from different sub ordinate level categories within in a general category. Fine grained classification is a very challenging task due to large inner-class variance and subtle inter-class distinctions. Variances in the pose, scale or rotation make the problem more difficult. In the recent years fine grained image classification has received considerable attention due to the advancement of deep learning based approaches. Layers of the features in deep learning techniques are not human designed instead learned from data using a general purpose learning procedure. There are large number of variants of deep learning architecture. In this survey we mainly focus on convolutional neural network (CNN) based approaches. Convolutional neural network(CNN) are surpassing other approaches in terms of accuracy and efficiency in a large margin. CNN is a type of feed-forward artificial neural network. It consists of one or more convolutional layers which are the building block of a CNN. The convolutional layers are then followed by one or more fully connected layers as in a standard multilayer perceptron(MLP).Most deep learning networks can be trained end to end efficiently using backpropogation.it is a common method of training artificial neural network used in conjunction with an optimization method such as gradient descent. In this survey ,we first introduce several convolutional neural networks which are mostly used for fine grained image categorization. Then part localization and object detection based approaches. The last section will review about feature extraction based approaches.

II. General deep network architectures

CNN is able to yield more discriminative representation of the image which is essential for fine grained image classification. AlexNet[1] is a deep convolutional neural network which won the ILSVRC-2012 competition with a top-5 test accuracy of 84.6% compared to 73.8% accuracy achieved by closest competitor. It consists of five convolutional layers, max pooling ones, Rectified Linear units(ReLus) as non-linearities, three fully connected layers. The Visual Geometry Group(VGG)[2] model has been introduced by visual Geometry Group from the university of Oxford.The VGG-16 has 13 convolutional Layers with 3 fully connected layers.VGG-19 has 3 more convolutional layers than VGG-16 model. Filters with a very small receptive field is used. All hidden layers are provided with the rectification non-linearity.

GooLeNet[3] is a network which won the ILSVRC-2014 challenge with a top-5 accuracy of 93.3%.This CNN is composed of 22 layers and a newly introduced building block called inception modules. The inception module allows for increasing the depth and width of the network while keeping computational budget constant. The 1x1 convolutions are used to compute reductions before the expensive 3x3 and 5x5 convolutions.

III. Part detection based approaches

A. *SPDA-CNN For Part Detection*

The semantic part detection and abstraction CNN architecture[4] uses two sub networks. One for detection and one for recognition. The detection sub-network uses a novel top-down method to generate small semantic part candidates for detection. The classification sub-network uses a novel part layers that extract features from parts detected by the detection sub-network and combine them for recognition. In detection sub-network k nearest neighbour(k-nn) method is used to generate proposals for semantic parts. Using k-nn, the detection network applies Fast R-CNN[] to regress and obtain much more accurate part bounding boxes. The final part detection are sent to abstraction and classification sub-network.

Based on the results from detection, the part Rol pooling layer does semantic pooling. This layer conducts feature selection and reordering which are useful for classification. By sharing the computation of convolutional filters, SPDA provides an end to end network that performs detection, localization of multiple semantic parts and whole object recognition within one framework.

B. *Deep LAC*

To recognize fine grained classes, the Deep LAC[5] incorporates part localization, alignment and classification in one deep neural network.it proposes Valve Linkage Function(VLF) for back-propagation chaining. The main network consists of three sub-networks for localization, alignment and classification. VLF connects all the sub networks and also function as information valve to compromise alignment and classification errors. For a given input image, the part localization sub network outputs the commonly used co-ordinates, top-left and bottom-right bounding box corners. Ground truth bounding boxes are generated with part annotations. This sub-network consists of 5 convolutional layers and 3 fully connected layers. The alignment sub-network receives part localization from localization sub-network.it performs template alignment. This network performs translation, scaling and rotation for pose aligned part region generation which is important for accurate classification. Apart from pose aligning, the VLF in the alignment sub-network plays a essential part in connecting the localization and classification sub networks. If the alignment is good enough in forward propagation, VLF guarantees accurate classification.

C. *Non-parametric transfer method*

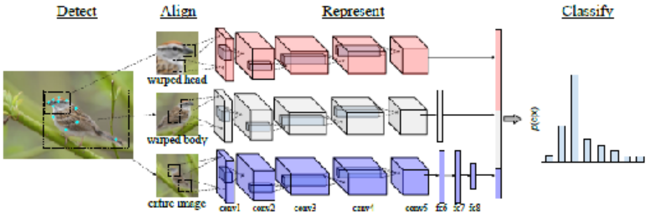
Non-parametric part transfer method[6] introduces a non-parametric approach for part detection.it is based on transferring part annotation from related training images to an unseen test image. The possibility for transferring part annotations to unseen images allows for copying with a high degree of pose and view variations. This method is valuable where intra-class variances are extremely high and precisely localized features need to be extracted. For training, a suitable feature representation is computed, which mainly focuses on the global shape of the object. The overall layout is represented with histograms of oriented gradients(HOG).During testing, again a global HOG feature is computed from the image region defined by the bounding box. Training images with an object shape similar to current test image is searched and then the part annotations are transferred from them directly. Final results are attained by pooling individual classification scores for every transferred part configuration. For classification SVM classifier is used.

D. *Part based R-CNN*

Part based R-CNN[7] learns the part detectors by leveraging deep convolutional features computed on bottom-up region proposal. R-CNN[8] is used to detect objects and localize their parts under geometric prior. Starting from bottom-up region proposals like selective search ,both object and part detectors are trained based on deep convolutional features. During testing all detectors are used to score, all the proposals and non-parametric geometric constraints are applied to rescore the proposals and choose the best object and part detection. Both the full objects bounding box annotations and a fixed set of semantic part annotations are used to train multiple detectors. A geometric constraints over the layout of the parts relative to the object localization is considered to filter the incorrect detections. At the final step, features for the whole object or part regions are extracted. Then SVM classifier is trained for the final categorization.

E. *Pose Normalized Convolutional Nets*

The pose normalized nets[9] first computes an estimate of objects pose and using this, it computes local image features. These local features are used for classification. The features are computed by applying deep convolutional nets to image patches which are located and normalized by the pose. The lower level feature layers are integrated with pose normalized extraction routines and higher level feature layers are integrated with unaligned image features.



Pose Normalized Convolutional Nets. This figure is from original paper[9]

In training, pose normalized nets uses Deformable Part Model(DPM)[10] to predict 2D locations and visibility of 13 semantic part key points. The classifier is trained with concatenated features, extracted from each prototype region and the entire image. In testing groups of key points are used to compute multiple warped image regions which are aligned using prototypical models. Each region is fed through a deep convolutional network. Features extracted are concatenated and fed to the classifier.

IV. OBJECT DETECTION BASED APPROACH

In this paper[11] two types of attention model is used using deep neural network for fine grained classification task. The raw candidate patches or image regions that have high objectness are generated in a bottom-up process. Selective search, an unsupervised process is used to extract patches from the input image. The bottom-up process will provide multi-view and multi-scale of the original image. Two top-down attention model is proposed in this paper. The object level top-down attention model filters the bottom-up raw patches. They remove the noisy patches that are not relevant to the object. This is done by converting a CNN pre-trained on ILSVRC2012 1K dataset into a FilterNet. A threshold score is used to decide whether a given patch should be selected.

The DomainNet, the second CNN is trained using the patches selected by the FilterNet. The DomainNet extracts features relevant to the categories belonging to the specific domain. It is a good fine grained classifier. Spectral Clustering is used to find the groups and use filters in a group to serve as part detector. Finally both the levels of attention model is merged to bring the significant gains. The advantage of this method is that the attention models are derived from CNN trained with classification task under weakest supervision setting where only class-labels are provided.

This Object-Part attention Model [12] has been proposed for weakly supervised fine grained image classification. The main novelties of object-part attention model are object level attention and part-level attention. The object level attention model localizes objects of images for learning object features and part level model is to select discriminative parts of object. Object level attention model based on saliency extraction, localizes the object of images automatically only with image level subcategory labels without any object or part annotation, both during testing and training. To filter out the noisy image patches, patch filtering is used. The saliency map is extracted via global average pooling in CNN, for localizing the objects of images.

V. FEATURE EXTRACTION

A. *Fused with Semantic Alignment*

Fused One-Vs-All Features(FOAF)[13] proposes a new framework for fine grained visual categorization. Semantic prior is combined with geometric information for efficient part localization. Using template based model, it detects less deformable parts and localizes other highly deformable parts with simple geometric alignment. During training and testing, no bounding boxes are provided. Less deformable part as well as object is first detected using R_CNN[8].A geometric refinement is done which restricts the detected head within the region of detected object. In order to segment foreground from background without object bounding boxes, object confidence map is computed, denoting the possible location of foreground object. The final alignment is based on the detected head and segmented foreground mask, which guarantees the accurate part alignment.

For Fused One-Vs-All Features learning, features are extracted based on the aligned parts and train One-vs-All SVM classifier to obtain mid-level features. Similar sub categories are fused iteratively and treated as micro-classes. Another SVM classifier is trained based on these mid-level features for final classification. FOAF achieves superior performance than traditional features for classification.

B. *Subset Feature learning*

Subset Feature learning[14] consists of two main parts. It proposes a subset learning system, which first clusters visually similar classes and then learns deep convolutional neural network for each subset. A progressive transfer learning system also been proposed to learn domain-generic convolutional feature extractor. Applying Linear Discriminant Analysis(LDA) to fc6 features to reduce their dimensionality, visually similar species are clustered into k subsets. A separate CNN is attained for each of the k pre-clustered subsets.

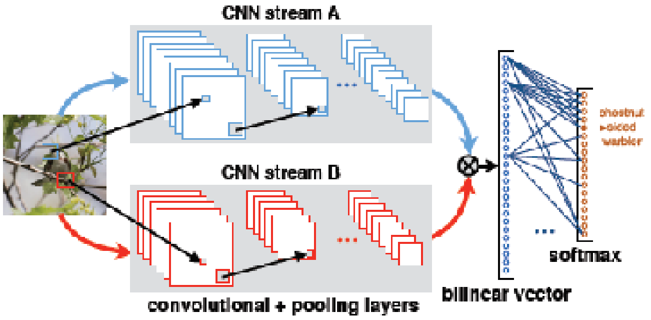
To select most relevant subset is to train a separate CNN based subset selector(SCNN).Using the output from the pre-clustering as the class labels, SCNN is trained by changing the softmax layer fc8 to make it have K outputs. As with the previously trained CNNs, the weights of SCNN are trained via back propagation.

C. *Selective Convolutional Descriptor*

In an unsupervised methodology of feature extraction Selective Convolutional Descriptor Aggregation(SCDA)[15] meet the challenges of automatically localizes the main object in fine grained images. A pretrained CNN model first extracts convolution activations for an input image. A Multiscale Orderless Pooling (MOP) CNN first extract CNN activations and perform VLAD pooling to determine the useful part of the activations. These useful descriptors are then aggregated using sum-pooled convolutional features on the last convolutional layer which outperforms the traditional methods like PCA(Principle Component Analysis). In SCDA, the preliminary step is getting pool5 which refers to the activations of the max-pooled last convolutional layer. To obtain a useful deep descriptors and removing background noise, aggregation map(A) is used. Then following this, mask map(M) of the same size as A is obtained. In order to resize it, bicubic interpolation is used whose size is same as input image.

D. *Bilinear CNN*

Bilinear Model[16] consists of two independent features combined with an outer product. A feature function takes as input, an image I and a location L and outputs a feature of size CxD. Generally locations can include position and scale. The feature outputs are combined with an outer product using the matrix outer product. To obtain an image descriptor, the pooling function aggregates the bilinear features across all locations in the image.



Bilinear CNN model. This figure is from original paper[16]

A natural candidate for the feature function f is a Convolutional neural network which consists of a hierarchy of convolutional and pooling layers. Two different CNNs pretrained on ImageNet datasets are used. By pretraining, bilinear deep network will benefit from additional training data in the cases of domain specific data scarcity. The variants of outer product representations are very effective at various texture, fine grained and scene recognition tasks.

E. *Refining Deep Convolutional Features*

Incorporating multi-scale image recognition using Fisher Vector(FV) CNN is proposed in Refining deep convolutional features[17].It is based on an unsupervised manner in extracting features. A non parametric feature weighting method is proposed on the basis of refining convolutional descriptors to boost the performance. To generate multiscale information pooling of feature tensors of the last convolutional layers with ReLu activation in a CNN is achieved using different pooling window sizes. No object bounding box and part annotations are used during training and testing times. Only image labels are used. Networks like VGG-16 and AlexNet are used

Table-1 Architecture and accuracy between different methods

Method	Architecture	Accuracy(%)
Part-based R-CNN	AlexNet	76.4
Pose Normalized Nets	AlexNet	75.7
Deep LAC	AlexNet	84.10
SPDA-CNN	VGGNET	85.14
Two-level attention Model	AlexNet	69.7
Bilinear CNN	VGGnet	77.2
Subset Feature Learning	AlexNet	77.5
Fused One Vs-All Feature	VGGNet	84.63

VI. CONCLUSION

This paper surveys some recent programs in deep learning based fine grained object detection and part localization. Several common convolutional neural networks are introduced first, including AlexNet, VGGNet and GooLeNet. They can be directly adapted to fine grained image classification. Since there are large intra class variance and small inter class variance in fine grained image classification, many common approaches resort to deep learning technology for object detection. While some approaches integrate the part localization into the deep learning framework, since subtle differences of visually similar fine grained objects exists in local parts. Some Fine grained categorization approaches also combine feature extraction to gain more classification capability and accuracy for fine grained images.

REFERENCES

[1]. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E.Howard, W. Hubbard, L. D. Jackel.,"Backpropagation applied to handwritten zip code recognition." *Neural Computation*,

[2]. K. Simonyan, A. Zisserman."Very deep convolutional networks for large-scale image recognition.", arXiv:1409.1556,2014

[3]. C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D.Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich,"Going deeper with convolutions." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 1–9, 2014.

[4]. H. Zhang *et al.*, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.

[5]. D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1666– 1674.

[6]. C. Göering, E. Rodner, A. Freytag, and J. Denzler, "Nonparametric part transfer for fine-grained recognition," in *Proc. IEEE Conf. Comput.Vis Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2489–2496.

[7]. N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Computer Vision— ECCV*. Cham, Switzerland: Springer, 2014, pp. 834–849.

[8]. W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural Network regularization," ArXiv:abs/1409.2329, 2014.

[9]. S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proc.BMVC*, Jun. 2014.

[10]. S. Branson, O. Beijbom and S. Belongie," Efficient large-scalestructured learning." in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Portland, USA, pp. 1806–1813, 2013.
.

[11]. T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep Convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 842–850.

[12]. Yuxin Peng , Xiangteng He, and Junjie Zhao, "Object-Part Attention Model for Fine-Grained Image Classification",*IEEE Trans. Image Process.*, vol. 27, no. 3,Mar. 2018.

[13]. X. Zhang, H. Xiong, W. Zhou, and Q. Tian, "Fused one-vs-all Features with semantic alignments for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 878–892, Feb. 2016.

[14]. Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Subset feature learning for fine-grained category classification," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 46–52.

[15]. X.S. Wei, J.H Luo and J.X. Wu, "Selective convolutional descriptor aggregation for fine-grained image retrieval", Xiv:1604.04994,2016

[16]. T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*,Dec. 2015, pp. 1449–1457.

[17]. Weixia Zhang,Jia Yan,Wenxuan Shi,Tianpeng Feng,Dexiang Deng" Refining deep convolutional features for improving fine-grained image recognition"