PARTICUL: Part Identification with Confidence measure using Unsupervised Learning

Romain Xu-Darme, Georges Quénot, Zakaria Chihani, Marie-Christine Rousset

PARTICUL: Part Identification with Confidence measure using Unsupervised Learning

Romain Xu-Darme[1,2], Georges Quénot[2], Zakaria Chihani[1], and Marie-Christine Rousset[2]

[1] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
{romain.xu-darme,zakaria.chihani}(at)cea.fr
{georges.quenot,marie-christine.rousset}(at)imag.fr

Abstract. In this paper, we present PARTICUL, a novel algorithm for unsupervised learning of part detectors from datasets used in fine-grained recognition. It exploits the macro-similarities of all images in the training set in order to mine for recurring patterns in the feature space of a pre-trained convolutional neural network. We propose new objective functions enforcing the locality and unicity of the detected parts. Additionally, we embed our detectors with a confidence measure based on correlation scores, allowing the system to estimate the visibility of each part. We apply our method on two public fine-grained datasets (Caltech-UCSD Bird 200 and Stanford Cars) and show that our detectors can consistently highlight parts of the object while providing a good measure of the confidence in their prediction. We also demonstrate that these detectors can be directly used to build part-based fine-grained classifiers that provide a good compromise between the transparency of prototype-based approaches and the performance of non-interpretable methods.

Keywords: Part detection · Unsupervised learning · Interpretability · Confidence measure · Fine-grained recognition

1 Introduction

With the development of deep learning in recent years, convolutional neural networks (CNNs) have quickly become the backbone of all state-of-the-art visual recognition systems. CNNs are highly efficient but also highly complex systems manipulating abstract (and often opaque) representations of the image - also called feature vectors - to achieve high accuracy, at the cost of transparency and interpretability of the decision-making process. In order to overcome these issues, a solution, explored in [1, 3, 6, 7, 10–12, 17, 19, 22, 27, 34, 41, 44], consists in building an intermediate representation associating each input image with a set of semantic *attributes*. These attributes, usually representing an association between a part of an object (*e.g.,* head of a bird) and a property (*e.g.,* shape, color), can be either used by interpretable methods [3,22,27] to produce the decision, as *post-hoc* explanation [11,12] of a particular decision, or as supplementary

information in order to discriminate similar categories in the case of fine-grained visual classification (FGVC) [7,10,19,41,44]. In practice, attributes are learned through fully supervised algorithms [1,6,7,10,19,34,41,44], using datasets with hand-crafted annotations. Such datasets are expensive to produce - using expert knowledge or online crowd-sourcing platforms - and prone to errors [13,23]. Therefore, a lot of effort [3,5,8,9,14,18,25,29,38,39,42,43] has been recently put into the development of more scalable techniques using less training information. In particular, the task of localizing different parts of an object, a prerequisite to attribute detection, can be performed in a weakly supervised (using the category of each image) or unsupervised manner. By focusing on general features of the objects rather than discriminative details, part detection represents an easier task than attribute detection. However, it must offer strong evidence of accuracy and reliability in order to constitute a solid basis for a trustworthy decision.

In this paper, we present PARTICUL[3] (Part Identification with Confidence measure using Unsupervised Learning), a plug-in module that uses an unsupervised algorithm in order to learn part detectors from FGVC datasets. It exploits the macro-similarities of all images in the training set in order to mine for recurring patterns in the feature space of a pre-trained CNN. We propose new objective functions enforcing the locality and unicity of the detected parts. Our detectors also provide a confidence measure based on correlation scores, allowing the system to estimate the visibility of each part. We apply our method on two public datasets, Caltech-UCSD Bird 200 (CUB-200) [35] and Stanford Cars [37], and show that our detectors can consistently highlight parts of the object while providing a good measure of the confidence in their prediction. Additionally, we provide classification results in order to showcase that classifiers based only on part detection can constitute a compromise between accuracy and transparency.

This paper is organized as follows: Section 2 presents the related work on part detection; Section 3 describes our PARTICUL model and Section 4 presents our results on two FGVC datasets. Finally, Section 5 concludes this paper and proposes several lines of research aiming at improving our approach.

2 Related work

Part detection (and more generally attribute detection) is a problem which has been extensively studied in recent years, especially for FGVC which is a notoriously hard computer vision task. Learning how to detect object parts in this context can be done either in a fully supervised (using ground-truth part locations), weakly supervised (using image labels only) or unsupervised manner.

Weakly supervised approaches [3,5,8,14,16,18,21,25,30,36,38,42,43] produce part detectors as by-products of image classification *i.e.,* object parts are jointly learned with categories in an end-to-end manner to help distinguish very similar categories. In particular, the OPAM approach presented in [25] obtains parts from candidate image patches that are generated by Selective Search [33], filtered

[3] Patent pending

through a dedicated pre-trained network, selected using an object-part spatial constraint model taking into account a coarse segmentation of the object at the image level, and finally semantically realigned by applying spectral clustering on their corresponding convolutional features.

Recently, unsupervised part detection methods have greatly benefited from the expressiveness and robustness of features extracted from images using deep CNNs. Modern approaches [15,29,39,40,42] mainly focus on applying clustering techniques in the feature space and/or identifying convolutional channels with consistent behaviors. More precisely, [15] produces part detectors through the sampling and clustering of keypoints within an estimation of the object segmentation produced by a method similar to GrabCut [26]. For a given image, [39] uses itemset mining to regroup convolutional channels based on their activation patterns, producing parts per image instead of globally (as in [25]) and thus requiring an additional semantic realignment - through clustering - for part-based classification. Rather than working on the full images, [40] and [29] use convolutional features of region proposals produced by Selective search [28]. [29] then learns the relation between regions and parts using clustering and soft assignment algorithms, while [40] builds part detectors from the top-k most activated filters across all regions of the training set. Due to the independent training of each detector, the latter approach produces tens of redundant part detectors which must be filtered out in a weakly unsupervised manner.

This issue is partially addressed in the MA-CNN approach [42], where part detectors are learned by performing a soft assignment of the convolutional channels of a pretrained CNN into groups that consistently have a peak response in the same neighborhood, then regressing the weight (importance) of each channel in each group through a fully convolutional layer applied on the feature map. Each part detector produces an activation map, normalized using a sigmoid function, corresponding to the probability of presence of the part at each location. By picking the location with the highest activation value for each detector, MA-CNN generates part-level images patches that are fed into a Part-CNN [2] architecture for classification. In practice, part detectors are initialized by channel grouping (using k-means clustering) and pre-trained in an unsupervised manner using dedicated loss functions enforcing the locality and unicity of each detector attention region. Finally, these detectors are fine-tuned during end-to-end learning of the image category. As such, this method falls into the category of unsupervised part learning with weakly supervised fine-tuning. More recently, the P-CNN approach [9] implements the part detection learning algorithm of MA-CNN, replacing the fully connected layer in charge of channel grouping by a convolution layer. Its classification pipeline uses Region-of-Interest (ROI) pooling layer around the location of maximum activation to directly extract part features and global features from the part detection backbone. In both cases, the channel grouping initialization requires first to process all images of the training set in order to cluster channels according to the location of their highest response. Moreover, for a given detector, only the area immediately surrounding the location of highest activation is taken into account (either to extract an

image patch in [42] or a vector through ROI-pooling in [9]), thus reducing the ability to detect the same part at different scales.

Contributions Our PARTICUL approach builds part detectors as weighted sums of convolutional channels extracted from a pre-trained CNN backbone. Our detectors are trained globally across the training set, instead of locally, contrary to [39] and [25] which require an additional semantic alignment phase. We develop objective functions that are specially crafted to ensure the compactness of the distribution of activation values for each part detector and the diversity of attention locations. Unlike [42] and [9], which propose similar functions, our approach does not require any fine-tuning of the backbone or an initial channel grouping phase. This induces a fast convergence during training and enables our module to be used with black-box backbones, in a plug-in fashion. Moreover, our detectors use a softmax normalization function, instead of sigmoid function, that simplifies the process of locating the part and enables us to detect parts at different scales. Finally, our detectors supply a measure of confidence in their decision based on the distribution of correlation scores across the training set. Importantly, this measure can also predict the *visibility* of a given part, a subject which is, to our knowledge, not tackled in any of the related work.

3 Proposed model

In this section, we present PARTICUL, our proposal for unsupervised part learning from FGVC datasets. In practice, each part detector can be seen as a function highlighting dedicated attention regions on the input image, in accordance with the constraints detailed in this section.

Notations For a tensor T inside a model $\mathcal{M}$, we denote $T(x)$ the value of T given an input x of $\mathcal{M}$. If T has H rows, W columns and C channels, we denote $T_{[h,w]}(x) \in \mathbb{R}^C$ the vector (or single feature) located at the h^{th} row, w^{th} column. We denote $*$ the convolution operation between tensors, σ the softmax normalization function. e define $\mathcal{I}$ as the set of all input images of a given dimension $H_{\mathcal{I}} \times W_{\mathcal{I}}$ and $\mathcal{I}_{\mathcal{C}} \subseteq \mathcal{I}$ as the subset of images in $\mathcal{I}$ containing an object of the macro-category $\mathcal{C}$. In our experiments, we denote $X_{train} = \{x_i \in \mathcal{I}_{\mathcal{C}} | i \in [1 \,..\, n]\}$ the training set of size n. Although $\mathcal{I}_{\mathcal{C}}$ cannot usually be formally specified, we assume that X_{train} is representative of this set.

3.1 Part detectors

As illustrated in Fig. 1, we first use a pre-trained CNN F to extract a $H \times W \times D$ feature map. Then, we build each of the p part detectors as a convolution with a $1 \times 1 \times D$ kernel $k^{(i)}$ (no bias is added), followed by a 2D spatial softmax layer, each part detector producing a $H \times W$ activation map

$$P(k^{(i)}, x) = \sigma\big(F(x) * k^{(i)}\big), \ \forall x \in \mathcal{I} \tag{1}$$

We denote the set of all part detector kernels as $K = [k^{(1)}, k^{(2)}, \ldots, k^{(p)}]$.

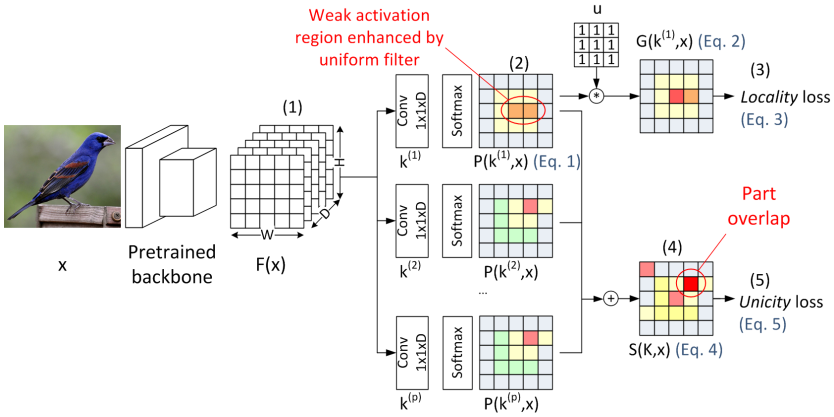


Fig. 1: Architecture of our PARTICUL model. (1) Convolutional features $F(x)$ are extracted from the image x. (2) Each part detector produces an activation map $P(k^{(i)}, x)$. (3) We apply a uniform kernel u to each part activation map before computing the Locality loss which ensures the compactness of activations. (4) All part activation maps are summed in $S(K, x)$. (5) Unicity loss is applied to ensure the diversity of part detectors. Here, part detectors 2 and p are very similar, leading to a high peak in bright red in $S(K, x)$. Best viewed in color.

3.2 Objective functions

Locality: For each part detector i, we force the network to learn a convolutional kernel $k^{(i)}$ which maximizes one region of the activation map for any training image x. In order to allow activations to be localized into a given neighborhood rather than in a single location in the $H \times W$ activation map, we first apply a 3×3 uniform filter u on $P(k^{(i)}, x)$. Let

$$G(k^{(i)}, x) = P(k^{(i)}, x) * \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{u} \tag{2}$$

Learning all convolutional kernels K in order to enforce locality can be translated as the optimization of the objective function

$$\mathcal{L}_l(K) = -\frac{1}{p} \sum_{i=1}^{p} \frac{1}{n} \sum_{x \in X_{train}} \max_{h,w} \left(G(k^{(i)}, x)_{[h,w]} \right) \tag{3}$$

Intuitively, solving the optimization problem in Eq. 3 using $G(k^{(i)}, x)$ rather than $P(k^{(i)}, x)$ relaxes the learning constraint and prevents the detectors from focusing on discriminative details between two adjacent feature vectors that would represent the same part.

Unicity In order to prevent the system from learning only a handful of easy parts, we wish to ensure that each feature vector $F_{[h,w]}(x)$ is not simultaneously correlated with multiple convolutional kernels $k^{(i)}$. Let $S(K,x) \in \mathbb{R}^{H \times W}$ s.t.:

$$S_{[h,w]}(K,x) = \sum_{i=1}^{p} P(k^{(i)},x)_{[h,w]} \tag{4}$$

Ensuring unicity can be translated as the following objective function, *i.e.,* making sure that no location (h,w) contains a cumulative activation higher than 1:

$$\mathcal{L}_u(K) = \frac{1}{n} \sum_{x \in X_{train}} \max\Big(\max_{h,w} S_{[h,w]}(K,x) - 1, 0\Big) \tag{5}$$

The final objective function for the unsupervised learning of part detectors is a weighted composition of the functions described above:

$$\mathcal{L}(K) = \mathcal{L}_l(K) + \lambda \mathcal{L}_u(K) \tag{6}$$

where λ controls the relative importance of each objective function.

3.3 Confidence measure and visibility

Visibility is related to a measure of confidence in the decision provided by each of our detectors and based on the distribution of correlation scores across the training set. After fitting our detectors to Eq. 6, we employ the function

$$\begin{aligned} H_i : \mathcal{I} &\rightarrow \mathbb{R} \\ x &\rightarrow \max_{h,w} \big(F_{[h,w]}(x) * k^{(i)}\big) \end{aligned} \tag{7}$$

returning the maximum correlation score of detector i for image x (before softmax normalization). The distribution of values taken by H_i across $\mathcal{I}_\mathcal{C}$ is modeled as a random variable following a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ estimated over X_{train}. We define the confidence measure of part detector i on image x as

$$C(x,i) = \Phi(H_i(x), \mu_i, \sigma_i^2) \tag{8}$$

where $\Phi(z, \mu, \sigma^2)$ is the cumulative distribution function of $\mathcal{N}(\mu_i, \sigma_i^2)$.

4 Experiments

In order to showcase the effectiveness of our approach, we apply our algorithm on two public FGVC datasets - the Caltech-UCSD Birds 200 (CUB-200) [35] dataset containing 11,788 images from 200 bird species (5994 training images, 5794 test images) and the Stanford Cars [37] dataset containing 16,185 images from 196 car models (8144 training images, 8041 test images). In addition to the object subcategory labels (bird species, car model), both datasets also provide additional information in the form of annotations (object bounding box, part locations in [35]). However, when learning and calibrating our part detectors, and unless specified otherwise, **we do not use any information other than the images themselves** and work in a fully unsupervised setting.

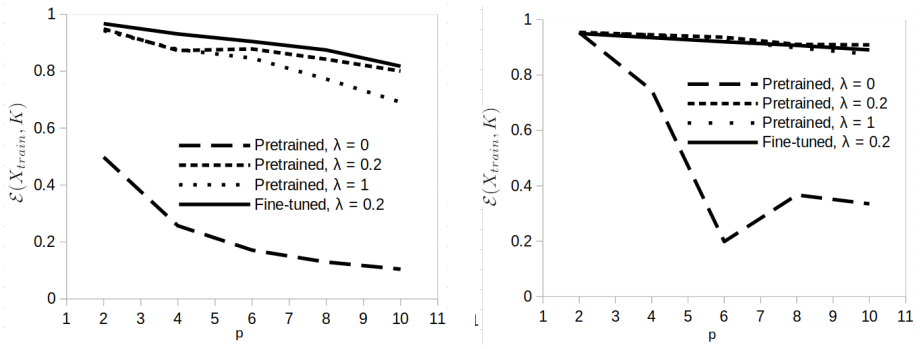
4.1 Unsupervised learning of part detectors

In this section, we illustrate our unsupervised part detection algorithm using VGG19 [31] (with batch normalization) pretrained on the Imagenet dataset [4] as our extractor $F(.)$ (Eq. 1). For part visualization, we use SmoothGrad [32], filtering out small gradients using the method described in [24].

We train our detectors during 30 epochs, using RMSprop with a learning rate of 5×10^{-4} and a decay of 10^{-5}. These training parameters are chosen using cross-validation on the training set, with the goal of minimizing the objective function (Eq. 6). Since we do not need to fine-tune the extractor, only $p\times D$ convolutional weights are learned during training, which drastically reduces the computation time. Finally, for a given number of parts p, we supervise the importance λ of the unicity constraint (Eq. 6) by measuring the overall attention across the feature map after training. More precisely, we compute the value

$$\mathcal{E}(X_{train}, K) = \frac{1}{|X_{train}|} \sum_{x \in X_{train}} \frac{1}{p} \sum_{h,w} \max_{i \in [1..p]} P(k^{(i)}, x)_{[h,w]} \tag{9}$$

corresponding to the average contribution of each detector. $\mathcal{E}(X_{train}, K) = 1$ corresponds to the ideal case where all detectors focus on different locations in the feature map $F(x)$, while $\mathcal{E}(X_{train}, K) = 1/p$ indicates a high redundancy of attention regions among detectors. As illustrated in Fig. 2, for both datasets



(a) CUB-200 (b) Stanford Cars

Fig. 2: $\mathcal{E}(X_{train}, K)$ v. number p of detectors for various values of λ, with pre-trained or fine-tuned extractor.

the average contribution of each detector decreases with p, reflecting the growing difficulty of finding distinct detectors. The choice of p itself depends on the downstream task (*e.g.,* classification), where again increasing the number of parts might produce diminishing returns (see Sec. 4.3). Moreover, in both datasets the presence ($\lambda > 0$) of our unicity loss function (Eq. 5) is paramount to learning distinct detectors. Without the unicity constraint, in the case of CUB-200 all

detectors systematically converge towards the same location inside of the feature map ($\mathcal{E}(X_{train}, K) \approx 1/p$) corresponding to the head of the bird; in the case of Stanford Cars, where images contain more varied distinctive patterns, the number of unique detectors depends on the random initialization of their convolutional weights, but the average contribution per detector quickly drops with p. For both datasets, we choose $\lambda = 0.2$ which maximizes $\mathcal{E}(X_{train}, K)$ for all values of p. As a comparison, we also train our detectors after fine-tuning the extractor on each dataset (in this case, the learning process can be considered weakly supervised), leading to only marginally better results. This supports our claim that, in practice, our detectors do not require a fine-tuned extractor and can work as a plug-in module to a black-box extractor.

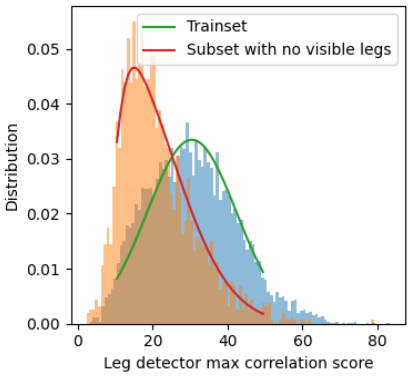
Fig. 3: Part visualization after training 6 detectors using the VGG19 extractor on Stanford cars (top) or CUB-200 (bottom). Using this visualizations, we can manually re-attach a semantic value to each detector, *e.g.,* the second part detector trained on CUB-200 is probably a "leg" detector. Best viewed in color.

As illustrated in Fig. 3, our detectors consistently highlight recurring parts of the objects and are relatively insensitive to the scale of the part. Although we

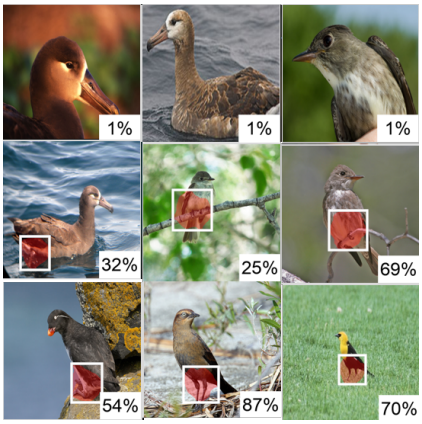
notice some apparent redundancy of the detected parts (*e.g.,* around the head of the bird), the relatively high corresponding value for $\mathcal{E}(X_{train}, K)$ (0.87 for $\lambda = 0.2$ and $p = 6$, see Fig. 2a) indicates that each detector actually focus on a different location of the feature map, leading to a richer representation of the object. Note that when training our detectors, we do not know beforehand towards which part each detector will converge. However, using part visualization, we can re-attach a semantic value to each detector after training.

4.2 Confidence measure

After training our detectors, we perform a calibration of their confidence measure using the method presented in Sec. 3.3. In order to illustrate the soundness of our approach, we exploit the annotations provided by the CUB-200 dataset to extract a subset of images where the legs of the bird are non-visible (2080 images from both the training and test set, where the annotations indicate that the color of the legs is not visible). As shown on Fig. 4a, there is a clear difference in



(a) Distribution of maximum correlation scores on the CUB-200 training set (in blue) and on a subset containing only images with non-visible legs (red).



(b) Confidence scores and part visualizations on images with non visible legs (top-row) and with visible legs (bottom rows).

Fig. 4: Confidence measure applied on a bird leg detector trained and calibrated on CUB-200 dataset. Best viewed in color.

the distributions of maximum correlation scores between images with and without visible legs. This also confirms that images not containing the part tend to produce lower correlation scores, and by consequence our proposal for a confidence measure based on a cumulative distribution function (Eq. 8). In practice (Fig. 4b), a calibrated detector can detect the same part at different scales while ignoring images where the confidence measure is below a given threshold (2% in

the example). It is interesting to note that for all images in the middle row, the annotations actually indicate that the legs are not visible, *i.e.,* our detectors can also be employed as a fast tool to verify manual annotations.

4.3 Classification

In all related works closest to ours, the performance of part detectors is never evaluated independently but rather *w.r.t.* to the classification task, a method which is highly dependent on the architecture chosen for the decision-making process: *e.g.,* SVMs in [15] and [29], Part-CNN [2] in [42], Spatially Weighted Fisher Vector CNN (SWFV) in [40], multi-stream architecture in [39]. Nevertheless, in order to provide a basis of comparison with state-of-art techniques, we also provide classification results based on the extraction of part feature vectors. As illustrated in Fig. 5, we use a method similar to [41], where the feature

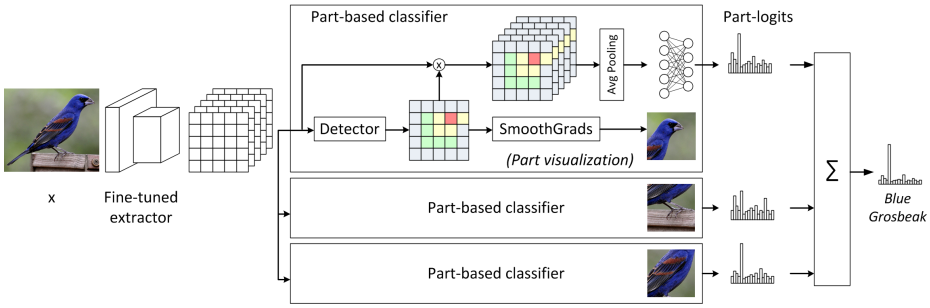


Fig. 5: Our classification model for FGVC datasets. The feature map produced by the extractor is masked out using the activation map of each detector and processed through a pooling layer, resulting in a set of part vector features. These vectors are processed independently through a set of fully connected layers to generate part-based logits that are summed up to produce the final prediction.

map $F(X)$ is multiplied element-wise with the activation map produced by each detector to compute a set of part feature vectors that are used as inputs of p independent classifiers (each containing 2 intermediate layers - with 4096 neurons, ReLU activation and dropout - before the classification layer). As such, our detectors operate a form of semantic realignment and extraction of relevant feature vectors from the image. Note that, contrary to the fixed size ROI pooling used in [9], this method has the advantage of dynamically adjusting the number of feature vectors taken into account depending on the scale of the part. For the final decision, we sum up the logits of all part-based classifiers to produce prediction scores for each category. Therefore, the final decision can be directly traced back to the individual result of each part-based classifier. This approach does not provide the transparency [20] of prototype-based methods [3,22], but constitutes a good compromise to non interpretable approaches using global features.

Table 1: Classification accuracy on CUB-200 and Stanford Cars and comparison with related works, from less transparent to most transparent.

Method	Train anno.	Accuracy (%) CUB-200	Accuracy (%) Stanford Cars
Global features only			
Baseline (VGG-19)		83.3	89.3
Part + global features			
UPM [39] (VGG19)		81.9	89.2
OPAM [25] (VGG-16)		85.8	92.2
MA-CNN [42] (VGG-19)		86.5	92.8
P-CNN [9] (VGG-19)		87.3	93.3
Part-based			
Ours (VGG-19)			
2 parts		70.1	76.0
4 parts		79.2	84.2
6 parts		81.5	87.5
8 parts		82.3	88.3
10 parts		82.3	88.6
OPAM [25] (parts-only, VGG-16)		80.7	84.3
No parts [15] (VGG19)	BBox	82.0	92.6
+ Test Bbox	BBox	82.8	92.8
PDFS [40] (VGG-19)		84.5	n/a
Prototypes			
ProtoPNet [3] (VGG-19)	BBox	78.0 ± 0.2	85.9 ± 0.2
ProtoTree [22] (Resnet-50)		82.2 ± 0.7	86.6 ± 0.2

← Transparency [20] →

Table 1 summarizes the results obtained when using our part-based classifier, along with the results obtained on a fine-tuned VGG-19 (acting as a baseline) and other state-of-the-art methods. With 8 detectors, we outperform the prototype-based approach [22] - which uses a more efficient extractor (Resnet50) - on Stanford Cars and obtain similar results on CUB-200. When compared with other methods using only part-level features, our PARTICUL model outperforms the OPAM [25] approach on both datasets. We also obtain comparable results to other methods requiring either the object bounding box [15] or to preselect detectors (in a weakly supervised manner) based on their classification accuracy [40]. When compared with less transparent methods using image-level (global) features in addition to part-level features - a method which usually has a significant impact on accuracy (*e.g.,* from 80.7% to 85.8% on CUB-200 using OPAM [25]) - again we achieve comparable results and even outperform the UPM [39] approach on CUB-200. Finally, it is also interesting to note that for both datasets, our classification results obtained by picking 10 feature vectors out of the $14 \times 14 = 196$ possible vectors of the extractor feature map correspond to a drop of less then 1% in accuracy when compared with the baseline, indicating that we are indeed selecting the most relevant vectors for the classification. Moreover, in 80% (resp. 77%) of these cases where only the baseline provides a correct prediction on the CUB-200 (resp. Stanford Cars) dataset, at least one of

our individual part-based classifier does provide a correct prediction (see Fig. 6). Thus, our proposed model could be further improved by fine-tuning the relative importance of part logits (*e.g.,* using the confidence measure).

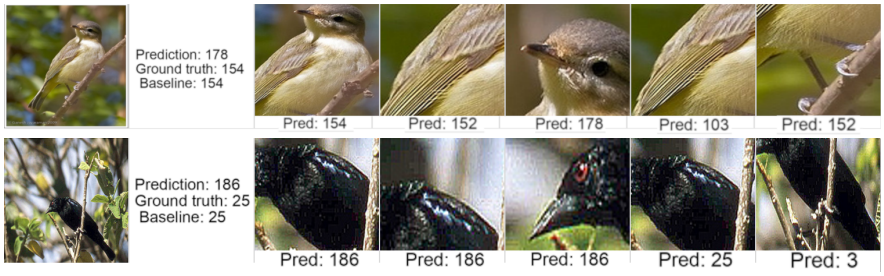


Fig. 6: Examples of images correctly classified by the baseline and incorrectly classified by our model (only 5 parts are shown for clarity). In both cases, at least one part-based classifier provides a correct prediction. Best viewed in color.

5 Conclusion and future work

In this paper, we presented our algorithm for unsupervised part learning using datasets for FGVC. We showed that our detectors can consistently highlight parts of an object while providing a confidence measure associated with the detection. To our knowledge, our method is the first to take the visibility of parts into account, paving the road for a solid attribute learning and ultimately for interpretable visual recognition. In the particular context of FGVC, our detectors can be integrated in a part-based classification architecture which constitutes a good compromise between the transparency of prototype-based approaches and the performance of non-interpretable methods. As a future work, we will study the integration of our detectors into a prototype-based architecture, learning prototypes from part feature vectors rather than from the entire image feature map. We will also study the impact of weighting part logits by the confidence score associated with the detected part on the overall accuracy of the system.

Acknowledgements Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations. This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003) and TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. This work was financially supported by European commission through the SAFAIR subproject of the project SPARTA which has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 830892, as well as through the CPS4EU project that has received funding from the ECSEL Joint Undertaking (JU) under GA No 826276.

References

1. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task CNN Model for Attribute Prediction. IEEE Transactions on Multimedia **17**(11), 1949–1959 (Nov 2015). https://doi.org/10.1109/TMM.2015.2477680, http://arxiv.org/abs/1601.00400, arXiv: 1601.00400
2. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. BMVC 2014 - Proceedings of the British Machine Vision Conference 2014 (Jun 2014), http://arxiv.org/abs/1406.2952, arXiv: 1406.2952
3. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: *This* looks like *That*: Deep learning for interpretable image recognition. Proceedings of the 33rd International Conference on Neural Information Processing Systems p. 8930–8941 (2019)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
5. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective Sparse Sampling for Fine-Grained Image Recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6598–6607. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00670, https://ieeexplore.ieee.org/document/9008286/
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing Objects by their Attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785 (2009). https://doi.org/10.1109/CVPR.2009.5206772
7. Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10697–10706 (2019). https://doi.org/10.1109/CVPR.2019.01096
8. Ge, W., Lin, X., Yu, Y.: Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3029–3038 (2019)
9. Han, J., Yao, X., Cheng, G., Feng, X., Xu, D.: P-cnn: Part-based convolutional neural networks for fine-grained visual categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 579–590 (2022). https://doi.org/10.1109/TPAMI.2019.2933510
10. Han, K., Guo, J., Zhang, C., Zhu, M.: Attribute-Aware Attention Model for Fine-grained Representation Learning. Proceedings of the 26th ACM international conference on Multimedia (2018)
11. Hassan, M.U., Mulhem, P., Pellerin, D., Quénot, G.: Explaining Visual Classification using Attributes. 2019 International Conference on Content-Based Multimedia Indexing (CBMI) pp. 1–6 (2019)
12. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating Visual Explanations. In: Computer Vision – ECCV 2016. pp. 3–19. Springer International Publishing (2016)
13. Jo, E.S., Gebru, T.: Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 306–316. FAT* '20, Association for Computing Machinery (2020). https://doi.org/10.1145/3351095.3372829, https://doi.org/10.1145/3351095.3372829

14. Korsch, D., Bodesheim, P., Denzler, J.: Classification-Specific Parts for Improving Fine-Grained Visual Categorization. German Conference on Pattern Recognition pp. 62–75 (10 2019)
15. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-Grained Recognition Without Part Annotations. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5546–5555 (2015). https://doi.org/10.1109/CVPR.2015.7299194
16. Kun Duan, Parikh, D., Crandall, D., Grauman, K.: Discovering Localized Attributes for Fine-Grained Recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3474–3481. IEEE, Providence, RI (Jun 2012). https://doi.org/10.1109/CVPR.2012.6248089, http://ieeexplore.ieee.org/document/6248089/
17. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-Based Classification for Zero-Shot Visual Object Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 453–465 (2014). https://doi.org/10.1109/TPAMI.2013.140
18. Li, H., Zhang, X., Tian, Q., Xiong, H.: Attribute Mix: Semantic Data Augmentation for Fine Grained Recognition. In: 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP). pp. 243–246 (2020). https://doi.org/10.1109/VCIP49819.2020.9301763
19. Liang, K., Chang, H., Shan, S., Chen, X.: A Unified Multiplicative Framework for Attribute Learning. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2506–2514. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.288, http://ieeexplore.ieee.org/document/7410645/
20. Lipton, Z.C.: The mythos of model interpretability. Communications of the ACM **61**, 36 – 43 (2018)
21. Liu, X., Xia, T., Wang, J., Yang, Y., Zhou, F., Lin, Y.: Fine-Grained Recognition with Automatic and Efficient Part Attention. [Preprint] arXiv:1603.06765 [cs] (Mar 2017), http://arxiv.org/abs/1603.06765
22. Nauta, M., van Bree, R., Seifert, C.: Neural Prototype Trees for Interpretable Fine-grained Image Recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14928–14938 (2021)
23. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. ArXiv **abs/2103.14749** (2021)
24. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
25. Peng, Y., He, X., Zhao, J.: Object-Part Attention Model for Fine-grained Image Classification. IEEE Transactions on Image Processing **27**(3), 1487–1500 (Mar 2018). https://doi.org/10.1109/TIP.2017.2774041, http://arxiv.org/abs/1704.01740, arXiv: 1704.01740
26. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut" — Interactive Foreground Extraction using Iterated Graph Cuts. ACM SIGGRAPH 2004 Papers (2004)
27. Rymarczyk, D., Struski, L., Tabor, J., Zieliński, B.: ProtoPShare: Prototype Sharing for Interpretable Image Classification and Similarity Discovery. [preprint] arXiv:2011.14340 [cs] (Nov 2020), http://arxiv.org/abs/2011.14340
28. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as Selective Search for Object Recognition. In: 2011 International Conference on Computer Vision. pp. 1879–1886. IEEE, Barcelona, Spain (Nov 2011). https://doi.org/10.1109/ICCV.2011.6126456, http://ieeexplore.ieee.org/document/6126456/

29. Sicre, R., Avrithis, Y., Kijak, E., Jurie, F.: Unsupervised Part Learning for Visual Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3116–3124. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.332, http://ieeexplore.ieee.org/document/8099815/
30. Simon, M., Rodner, E.: Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 1143–1151 (2015)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2015)
32. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. ArXiv **abs/1706.03825** (2017)
33. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. International Journal of Computer Vision **104**, 154–171 (2013)
34. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute Recognition by Joint Recurrent Learning of Context and Correlation. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 531–540 (2017)
35. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
36. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 842–850 (2015)
37. Yang, L., Luo, P., Loy, C.C., Tang, X.: A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3973–3981 (2015)
38. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to Navigate for Fine-grained Classification. In: Computer Vision – ECCV 2018. pp. 438–454. Springer International Publishing (2018)
39. Zhang, J., Zhang, R., Huang, Y., Zou, Q.: Unsupervised Part Mining for Fine-grained Image Classification. [preprint] ArXiv **abs/1902.09941** (2019)
40. Zhang, X., Xiong, H., gang Zhou, W., Lin, W., Tian, Q.: Picking Deep Filter Responses for Fine-Grained Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1134–1142 (2016)
41. Zhao, X., Yang, Y., Zhou, F., Tan, X., Yuan, Y., Bao, Y., Wu, Y.: Recognizing Part Attributes With Insufficient Data. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 350–360. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00044, https://ieeexplore.ieee.org/document/9009781/
42. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5219–5227. IEEE, Venice (Oct 2017). https://doi.org/10.1109/ICCV.2017.557, http://ieeexplore.ieee.org/document/8237819/
43. Zhou, X., Yin, J., Tsang, I.W.H., Wang, C.: Human-Understandable Decision Making for Visual Recognition. In: PAKDD (2021)
44. Zhu, J., Liao, S., Lei, Z., Li, S.: Multi-label Convolutional Neural Network Based Pedestrian Attribute Classification. Image and Vision Computing **58**, 224–229 (2017)