I Have Seen Enough: Transferring Parts Across Categories

David Novotny[1]
david@robots.ox.ac.uk
Diane Larlus[2]
diane.larlus@xrce.xerox.com
Andrea Vedaldi[1]
http://www.robots.ox.ac.uk/~vedaldi

[1] Visual Geometry Group
University of Oxford
[2] Computer Vision Group
Xerox Research Centre Europe

Abstract

The recent successes of deep learning have been possible due to the availability of increasingly large quantities of annotated data. A natural question, therefore, is whether further progress can be indefinitely sustained by annotating more data, or whether there is a saturation point beyond which a problem is essentially solved, or the capacity of a model is saturated. In this paper we examine this question from the viewpoint of learning shareable semantic parts, a fundamental building block to generalize visual knowledge between object categories. We ask two research questions often neglected: whether semantic parts are also visually shareable between classes, and how many annotations are required to learn them. In order to answer such questions, we collect 15,000 images of 100 animal classes and annotate them with parts. We then thoroughly test active learning and domain adaptation techniques to generalize to unseen classes parts that are learned from a limited number of classes and example images. Our experiments show that, for a majority of the classes, part annotations transfer well, and that performance reaches 98% of the accuracy of the fully annotated scenario by providing only a few thousand examples.

1 Introduction

Recent progress in image understanding, while dramatic, has been primarily fueled by the availability of increasingly large quantities of labeled data. For example, it was only with the introduction of resources such as ImageNet that deep learning methods were finally able to realize their potential. However, it is unclear whether manual supervision will be able to keep up with the demands of increasingly sophisticated and data-hungry algorithms. New projects such as the Visual Genome project [26], where millions of image regions are labeled with short sentences and supporting bounding boxes, go well beyond standard datasets such as ImageNet and offer new terrific research opportunities. At the same time, however, they raise the obvious question: *when is supervision enough?*

The idea that limitless manual supervision is impractical has motivated research in areas such as unsupervised learning or learning from off-the-shelf resources such as the Web. While these are important directions, such approaches go to the other extreme of avoiding

Pages 115.1-115.14
DOI: https://dx.doi.org/10.5244/C.30.115

Figure 1: *ImageNet Animal Parts.* We investigate the ability of semantic parts to transfer between different categories. To do so, we extend 100 ImageNet animal categories with selected part annotations (the figure shows one annotated example per class). We then ask the question: what is the minimal level of supervision so that part detection trained on the source classes perform well on the target classes?

manual supervision altogether. In this paper, we take a pragmatic approach and start from the assumption that explicit manual supervision is currently the most effective way of constructing models. However, we also ask whether there is a limit to the amount of supervision which is actually required and hence of data that needs to be annotated.

While answering this question in full generality is difficult, we can still conduct a thorough analysis in representative cases of general interest. In this paper, we look in particular at the problem of *recognizing and detecting semantic parts in categories* (*e.g.* animal eyes; Fig. 1), because semantic parts are highly informative, and, importantly, *semantically shareable* (*e.g.* both monkeys and snakes have eyes which, despite important differences, are broadly analogous). In fact, one of the key uses of parts in computer vision is to transfer the structure of known objects to novel ones, for which no supervision is available. However, an important practical question, which is often neglected, is whether semantically shareable parts are also *visually shareable*, in the sense of being recognizable in novel objects without further supervision. This assumption has never been tested before beyond a few categories or narrow domains.

In this paper, we conduct the first careful investigation of *part transferability across a large target set of visually dissimilar classes*. Since there is no dataset suitable for this analysis, our first contribution is to augment a subset of the ImageNet ILSVRC data [8] consisting of 100 animal classes with annotations for selected semantic parts[1] (section 2).

Given the new data, we investigate the two key aspects of transferability under bounded supervision (section 3): (i) learning parts from a limited number of example images and (ii) applying known parts to new, unseen domains. For the first problem, we consider an *active learning* scenario, where images are annotated with parts in turn, studying which images should be chosen and how many are needed to saturate performance.

Next, we look at part transferability as a *Domain Adaptation* (DA) problem. Differently from the typical transductive learning scenario of DA, where algorithms leverage unlabeled data of the target domain to adapt a source predictor, our goal is to learn part predictors that apply directly to novel classes *before having samples from those*. We propose to address this domain generalization problem using an efficient ensemble of detectors which is optimized for generalization to new classes and that, at the same time, can be used to guide active learning.

Finally, we conduct a thorough empirical evaluation of these problems (section 4). As we do so, we provide insights on how subsets of images may be selected for labeling, and how many such labels may be required to perform well on unseen categories.

1.1 Related work

Our work relates to several research topics including domain adaptation and active learning. **Domain adaptation** (DA) is a special case of transfer learning that seeks to learn predictors that account for a shift between the distributions of the source and target data domains. Typically, only the source domain is labeled. Since the seminal paper of Saenko *et al*. [41], DA has been applied to computer vision by learning feature transformations [2, 13, 17, 18, 46], or by adapting the parameters of the predictor [21, 60]. Only a few papers consider DA from more than one source domain [47], and most from at most a few, while we consider 50. Sometimes, the source domains are not given a priori, but discovered implicitly [20]. Recently, DA has also been applied to deep learning: the works of [15, 53] learn domain-invariant features by defeating a classifier that tries to separate the source and target domains; [52] improves the robustness to domain shift of the final layer of a CNN.

All these approaches formulate DA as *transductive* learning, for which they require unlabeled samples from the target domain. This is a fundamental difference from our case, where *no target samples are available*, also known as domain generalization [14, 16, 34] or predictive domain adaptation [62].

Active learning. The goal of active learning is to reduce the annotation costs by deciding which training samples should be annotated. Each annotation is associated with a cost [6] and the goal is to obtain the best performance within a budget. Many data selection strategies have been proposed, based on uncertainty and entropy [48], used in [7, 23, 25, 36], or diversity and representativeness [24]. The work of [56] estimates a cost of different types of annotations and then modulates an expected risk function while the strategy of [40] annotates as many examples in each image as possible. [35] leverages additional information from different annotation types to improve the convergence properties of the active learner. Only very few works have jointly looked at transfer learning and active learning [3, 37, 42, 58, 59]

[1] The dataset can be downloaded at `www.robots.ox.ac.uk/~vgg/data/animal_parts`.

as we do here, and none of them considered computer vision tasks. Moreover, the transfer learning components of these works approach the transductive domain adaptation task whereas we focus on domain generalization.

Related transfer learning problems. Zero-shot learning, *i.e.* the task of recognizing a category with no training samples, is often tackled by explicitly learning classifiers that are transversal to object classes. This can be done by modeling semantic relatedness [38, 39], or by transferring other knowledge such as materials [5, 54], segments [45], parts [10, 49] or attributes [11, 29]. However, these works consider only a small number of classes and *assume* that primitives such as parts transfer visually, whereas here we explicitly question this assumption. Fewer works consider transfer learning for localization as we do; these include the works of [9, 19, 28, 33] that transfer bounding box information using the ImageNet hierarchies; the method of [22] that transfer object detectors from seed classes; and [1] which transfers detectors assuming a limited number of annotated examples in the target domain. Differently from such works, we do not transfer whole objects, but individual keypoints, and we do so between very diverse classes. Transferring keypoints was explored in [63], which detects facial landmarks using a deep multi-task learning framework, while [51] induce pose for previously unseen categories.

2 A new dataset to study semantic parts

A thorough evaluation of the transferability of parts requires a suitable dataset with a large enough number of classes. Unfortunately, datasets that have keypoints or part-level annotations either consider a handful of classes, such as the PASCAL Parts [4], or are specialized to a narrow set of domains, focusing only on birds [57], faces [32, 63], or planes [55]. Datasets with more categories, such as the Visual Genome [26], do not contain systematic part annotations. Instead of collecting a new set of images, we build on top of the existing ImageNet dataset [8]. In addition to being a familiar dataset, its classes are already organized in a semantic hierarchy, induced by WordNet [12]. This provides a natural basis to study the semantic structure of this space. A very significant challenge with annotating many parts for many object categories is of course the very large cost; thus, trade-offs must be made.

Here, the singly most important aspect for experimentation is to label a sufficiently large space of categories. This space should also contain a mix of similar and dissimilar categories. Furthermore, the same parts should ideally apply to all categories. Here we select the "vertebrate" subtree of the ImageNet ILSVRC. This tree contains 233 animal classes, of which we select 100 for experimentation (Figure 1). We annotate two parts. One, *eyes*, exist in all selected animals. The second, *feet*, exist in a large subset of these (mammals and reptiles but not fish). Beyond their semantic shareability (visual shareability has to be verified), these parts were selected because they are easily understood by annotators from crowdsourcing platforms, and they can satisfactorily be annotated with keypoints as opposed than by drawing bounding boxes or regions. Both properties were instrumental in collecting a large dataset of part annotations in a reasonable time and within a reasonable budget. While limited, these annotations are sufficient to demonstrate the principles of our analysis. We collected annotations for about 150 images per class, annotating 14711 images in total.

3 Methods

First, we describe the keypoint detector that we use to detect semantic parts in images. Second, as our experiments sample images in turn, we describe how uncertainty sampling can be defined in our particular active learning scenario. Finally, we describe how the fact that we have many source domains (*i.e.* animal classes) to sample from can be leveraged in order to combine several detectors that are then applied to the unseen target classes.

Keypoint detector. As our baseline keypoint detector, we use the state-of-the-art method proposed by Tulsiani and Malik [50]. This architecture uses convolutional layers from VGG-VD [44], followed by a linear regressor, that outputs a heatmap expressing the probability of a keypoint being centered at a particular location. The accuracy is improved by linearly combining the outputs of a coarse-scale (6×6) and a fine-scale (12×12) network.

As our analysis requires frequent retraining of the model, we adapt the faster 6×6 network of [50] to output finer-scale 12×12 cell predictions. To do so, we follow [30] and we append a bilinear upsampling layer to the final keypoint regressor convolutional layer and linearly combine the resulting upsampled heatmap with a finer-scale convolutional heatmap derived from the pool4 VGG-VD layer. The recombined heatmap is terminated with a sigmoid activation function. The resulting architecture allows for multiscale end-to-end training while increasing overall testing/training speed by a factor of 3.

Active learning. The goal of active learning is to select a small subset of images for annotation while maximizing the performance of the final predictor. Let U be the set of all available images; the algorithm starts by a pool $L_0 \subset U$ containing $|L_0| = 50$ randomly-selected images $x \in U$ and collects the corresponding annotations. Then, for $t = 0, 1, 2, \ldots$ the algorithm alternates training a CNN keypoint detector using the annotations in L_t and collecting annotations for A more images. For the latter, all non-annotated images in $x \in U$ are assessed and the A "most informative ones" are selected for annotation.

The standard criterion to select informative images is to pick the ones which leave the current predictor uncertain, also called *uncertainty sampling*. However, while uncertainty is easily defined in classification tasks where the goal is to predict a single label per image, it is not obvious how to do so for keypoint prediction where the predictor produces a score for each image location. We propose to do so as follows: let $p(y = +1|x,u) = \Phi(x)_u$ be the probability of finding a keypoint at location u in image x as computed by the CNN Φ (unless otherwise specified, we assume that the CNN is terminated by a sigmoid function). The uncertainty score is then given by $1 - 2|\max_u p(y = +1|x,u) - 1/2|$. Intuitively, when the model is certain, either (i) there are no keypoints in that image and $\max_u p(y = +1|x,u) \approx 0$, *or* (ii) there is at least one keypoint and then $\max_u p(y = +1|x,u) \approx 1$.

Transfer learning by auto-validation. Our problem differs from standard active learning in that, as in DA, the target classes are *shifted* compared to the source ones. Furthermore, differently from the standard transductive learning setting in DA, our aim is to learn a "universal" part detector that generalizes to unseen target classes without further training.

Compared to more common machine learning settings, we can leverage the fact that the source data is split in well defined domains to characterize domain shift and improve generalization. We do so using an *auto-validation* procedure. Let $D = \{d_1, \ldots d_N\}$ be a set of source domains (object classes) and let $\delta \subset D$ a subset of the latter. We train CNN part predictors Φ_δ for each such subset and recombine them to maximize generalization using the method of Krogh *et al.* [27]. Recombination requires computing the cross-validation error E_δ of model Φ_δ on the complementary domains $D - \delta$, as well as the cross-correlation

matrix $C_{\delta\delta'} = E_{xu}[\Phi_\delta(x)_u \Phi_{\delta'}(x)_u]$ of the response (heat maps) of the different CNNs on the training data. Given E_δ and C, the combined detector is given by $\bar{\Phi} = \sum_{\delta \subset D} \alpha_\delta \Phi_\delta$ where the coefficients α_δ are obtained by solving the following quadratic programming problem [27]:

$$\begin{aligned} \underset{\{\alpha_\delta | \delta \subset D\}}{\arg\min} \quad & \sum_{\delta \subset D} \alpha_\delta E_\delta + \sum_{(\delta,\delta') \subset D \times D} \alpha_\delta C_{\delta\delta'} \alpha_{\delta'} - \sum_{\delta \subset D} \alpha_\delta C_{\delta\delta} \\ \text{s.t.:} \quad & \alpha_\delta \geq 0, \ \forall \delta \subset D \\ & \sum_{\delta \subset D} \alpha_\delta = 1 \end{aligned} \tag{1}$$

The original method from [27] was designed to recombine predictions of independently trained shallow neural networks. We adapt the original method so the different keypoint detectors Φ_δ share weights for the early layers, thus removing the requirement of costly re-optimization of the large number of CNN parameters on each set of training samples δ. In practice, we propose to decompose the CNN as $\Phi_\delta = \phi_\delta \circ \phi_0$, where ϕ_0 is the same for different δ and only ϕ_δ is specific to δ. More precisely, ϕ_0 includes all learnable parameters of the original VGG-VD layers up to conv5_3 and ϕ_δ comprises the final convolutional filter terminated by the sigmoid layer that outputs part detector responses specific to training samples from δ.

Optimization of ϕ_δ and ϕ_0 is easily implemented using stochastic gradient descent (SGD): given a data sample x, for all δ in parallel, either ϕ_δ or the cross-validation error E_δ are updated, depending on whether $x \in \delta$. The cross-correlation matrix C is estimated using all samples irrespective of their origin. To ensure numerical stability we add a small constant λ to the diagonal of C ($\lambda = 0.1$ in all experiments). Once optimization of ϕ_δ, E_δ and C completes, coefficients α_δ are obtained by solving eq. (1).

Another advantage of training an ensemble of detectors Φ_δ is that their lack of agreement on the training data can replace uncertainty sampling in guiding active learning. We implement this query-by-committee [43] criterion (QBC) following [27]: Given a pixel u in a test image x we assess the disagreement between pixel-wise predictors $\Phi_\delta(x)_u$ by evaluating the *ensemble ambiguity* $a(x,u) = \sum_{\delta \subset D} \alpha_\delta (\Phi_\delta(x)_u - \bar{\Phi}(x)_u)^2$, where $\bar{\Phi}(x)_u = \sum_{\delta \subset D} \alpha_\delta \Phi_\delta(x)_u$. Similar to uncertainty sampling (section 3) we label each image x with a disagreement score $\hat{a}(x)$ by max-pooling over the pixel-wise ensemble ambiguities, i.e. $\hat{a}(x) = \max_u a(x,u)$. During the labeling stage of active learning, samples with highest $\hat{a}(x)$ are added first.

4 Experiments

In this section we first perform a quantitative evaluation of part transferability (section 4.1) followed by evaluation of the proposed active-transfer learning methods (section 4.2).

Experimental protocol. The set of 100 domains (animal classes) is split into 50 source domains and 50 target domains as follows. To achieve uniform coverage of the animal classes in both sets, we first cluster the 100 classes into $K = 50$ clusters using their semantic distance and spectral clustering. The semantic distance between two classes d and d' is defined as $|r \to d \cap r \to d'| / \max\{|r \to d|, |r \to d'|\}$, where $|r \to d|$ is the length of the path from the root of the hierarchy to class d. Then, each cluster representative is included in the target set and the complement is included in the source set. Furthermore, images in each class are divided into a 70/30 training-testing split, resulting in four image sets: source-train, source-test, target-train and target-test. As common practice [31, 61], keypoint detections are restricted to ground-truth object bounding boxes for all evaluation measures.

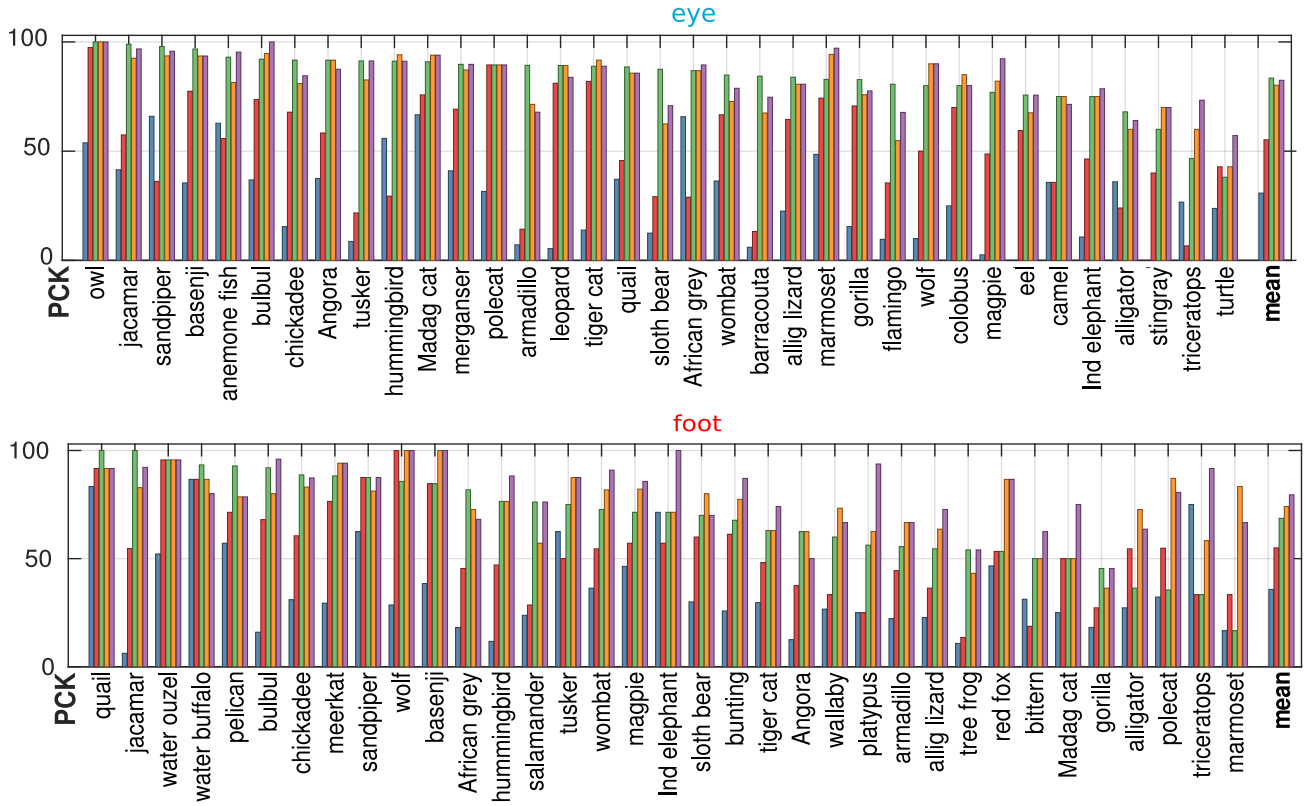

Figure 2: **Relative difficulty of part detection and part transfer.** Part detection performance for eyes (top) and feet (bottom) for a subset of the target classes, where the detector has been trained using either: the farthest class (in semantic distance), the nearest class, the same class, the source classes, all source and target classes. Classes are sorted by increasing difficulty.

Evaluation measures. We evaluate keypoint detection using two standard metrics [61]: PCK and APK. In **PCK**, for each ground truth bounding box, an algorithm predicts the single most confident keypoint detection. This detection is regarded as true positive if it lies within $\alpha \max\{w,h\}$ of the nearest ground truth keypoint, where w,h are the box dimensions and $\alpha \in \langle 0,1 \rangle$ controls the sensitivity of evaluation to misalignments. For **APK**, keypoints are labeled as positive or negative detections using the same criterion as PCK and ranked by decreasing detection scores to compute average precision. In all of our experiments we set $\alpha = 0.05$ for the eyes, that are small and localized, and $\alpha = 0.1$ for the feet which are more difficult to annotate with a keypoint.

Baseline detector. We validated our faster baseline by comparing it to the original $6 \times 6 + 12 \times 12$ model of [50]. Our implementation of the $6 \times 6 + 12 \times 12$ architecture achieves 61.1% PCK on the PASCAL VOC rigid keypoint detection task – a comparable result to 61.5% PCK reported in [50]. We also experimented on our dataset, and results show that our 6×6 upsample architecture is very competitive while being much faster than alternatives.

4.1 Visual shareability of parts

In this section we challenge the idea that parts are visually "shareable" across different classes and that, therefore, it suffices to learn them from a limited number of classes to understand them equally well in all cases. Figure 2 shows part detection performance for individual classes, for different configurations that we discuss below.

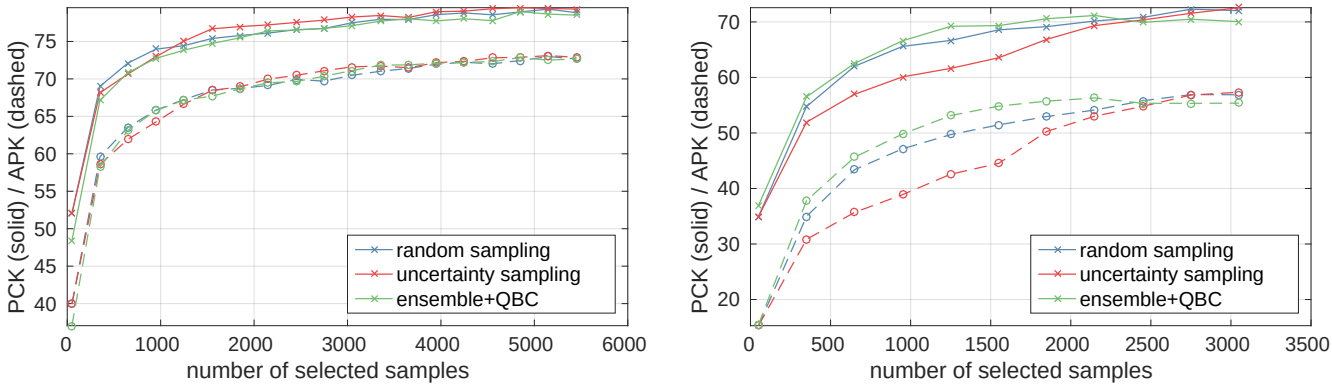

Figure 3: **Active-transfer learning**. PCK (solid lines) and APK (dashed lines) as a function of the number of used labeled examples for random sampling (RS), uncertainty sampling (US) and the network ensemble with query-by-committee sampling (ensemble+QBC) methods, for the eye (left) and foot (right) parts.

Learning from a single class. We first look at the individual target class detection results when learning from annotated samples from the same class (green bar plots). We see that the difficulty of detecting a certain part strongly depends on the specific class. For example, owl's eyes have 100% PCK, whereas turtle's eyes have 38.1 PCK. We then compare with two other training sets of identical size: i) with the nearest class (NC - red bar plot) according to the semantic measure, and ii) with the farthest class (FC - blue bars). As expected, we verify that NC outperforms FC by 20.9% PCK in average (NC is better than FC in 39 classes out of 50 for the eyes, and 32 out of 37 for the feet). This demonstrates the relevance of the semantic distance for cross-domain transfer. In average NC still performs 26.9% below training with the target class itself. Next, we consider transferring from more classes.

Increasing the training set. We compare the performance of detectors when these are trained with larger subsets of the data: i) using all classes available (*i.e.* the source and target domains, purple bar plots), and ii) using only the source domains (that do not contain the target class, orange bars). We note several facts. First we observe that using all classes improves compared to training only for the target class in average for feet, but not for eyes that perform very well already. Then, we observe that in 61% of the cases, learning a part from source classes alone or adding the target classes changes PCK by less than 7%. Hence, if parts are learned form a sufficiently-diverse set of classes, they can be expected to transfer satisfactorily to novel ones as well. In average, training from the source classes only (transfer scenario) is only 2.2 PCK below training from the full set of classes for eyes, and only 5.5 PCK below for feet.

4.2 Active-transfer learning

In the previous section we looked at how well parts transfer from known (source) classes to new (target) classes. Here we ask how many source images need to be annotated in the source domain. In order to answer this question, we adopt the active-transfer learning framework of section 3 and we monitor the performance of the detector on the target classes as more annotated images for the source classes become available. The overall performance is summarized by plotting the attained mean PCK (solid lines) and APK (dashed lines) as a function the number of labeled source images (fig. 3). We compare three methods: active learning by random sampling (RS), active learning by uncertainty sampling (US), and network ensemble with active learning by query-by-committee (ensemble+QBC).

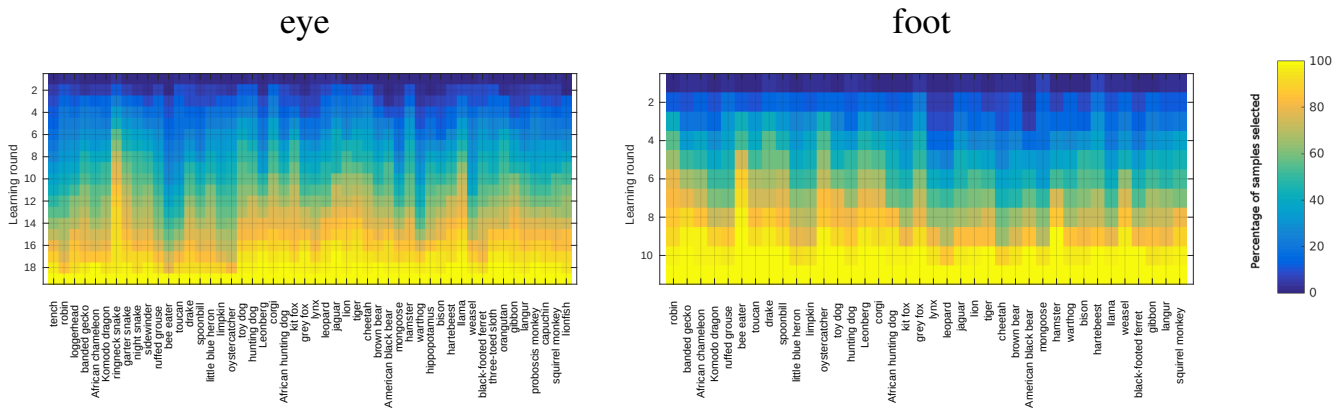

Figure 4: **Comparative domain importance.** The figure shows, for each of the source classes, how many more images active learning selects at each round by ensemble+QBC. Left: eye detection; Right: foot detection.

Implementation details. The initial pool contains $|L_0| = 50$ randomly-selected image annotations and further 300 image annotations are added at every active learning round. For pretraining the CNN we first remove all the images of the vertebrate subtree from the original ILSVRC12 dataset and then train according to the protocol from [44]. In each active learning round, the CNN from the previous round is fine-tuned for 7 epochs, lowering the learning rate tenfold after epoch 5 (this was verified to be sufficient for convergence). During early stages of active learning, duplicate training samples are added to ensure that at least 300 learning iterations are performed per epoch. Learning uses SGD with momentum and mini-batch size of 20. Mini-batches are sampled to guarantee that all the training classes are equally represented on average (rebalancing). Momentum, weight decay, and initial learning rate were set to 0.9, 0.005, and 0.0003 respectively. Training sets are augmented by jittering the sizes of image crops and by adding their horizontal flips. All parameters were validated by splitting the source domains in half. For DA by auto-validation the set D of possible source domains was obtained by regrouping the 50 source domains into 3 super-domains by clustering them using their semantic similarity. The resulting ensemble contains 6 experts trained and auto-validated on pairs of complementary domains $(\delta, D-\delta)$ and an additional expert that is both trained and auto-validated on all source domains. All experiments are repeated 4 times with different random seeds and averages are reported.

Results. First, we observe that US performs slightly better than the other two algorithms on the eye, but is substantially outperformed by RS and ensemble+QBC on the feet class. We found that the reason is that the network is typically most uncertain about images that happen to not contain any part instance, which is fairly frequent with animal feet as they tend to be occluded. On the contrary, RS is not affected by this problem. Ensemble+QBC performs as well as RS on the eye part and noticeably better on the foot part. This indicates that guiding active learning using the QBC criterion is more robust than US. The fact that the ensemble+QBC method performs similarly to the others on the eye class is likely due to the fact that there is less visual variability in eyes than feet and therefore all classifiers in the ensemble are similar, with poorer generalization [27]. Ensemble+QBC also benefits from improved generalization by the optimized ensemble of domain-specific models. Finally, we verified that using the ensemble of models with uncertainty sampling strategy is still not competitive. We conclude that ensemble+QBC is an effective active-transfer learning strategy.

Besides the relative merits of individual active learning strategies, a main observation

for our investigation is how quickly performance of different methods saturates. It can be noticed that for eyes the performance reaches 2% of the maximum with around 3,000 annotations, and for feet, the performance reaches 2% of the maximum with around 2,100 annotations. Combined with the observations in section 4.1, this indicates that excellent performance can be achieved for part detection in most animal classes by annotating a small representative subset of classes and a small number of corresponding images. This result is somewhat remarkable and can be attributed to the excellent performance of pre-trained deep neural networks as general-purpose representations. Recall that networks were pre-trained for image classification and not part detection, and not on any of the source or target classes.

Sampling strategy analysis. Figure 4 shows the distribution of selected animal classes during individual learning rounds for the QBC strategy. The distribution is clearly non-uniform, and the method seems to select representative classes within groups such as "reptiles", "felines", etc.

5 Conclusions

In this paper we have looked at the problem of semantic part transferability in image understanding. Semantic parts are often assumed to be a good vehicle for generalization, but this hypothesis has seldom been tested explicitly using a large number of different classes. We have done so by creating a new dataset of annotated parts in the ImageNet ILSVRC 2012. Then, we have looked at two main questions: whether parts trained on a set of representative classes generalize to others and how many images are required to train such classes using state-of-the-art neural network detectors and methods for active-transfer learning.

Our main finding is that parts transfer well to the majority of new classes even if trained from a limited number of examples. This is a very encouraging result that suggests that the underlying pre-trained deep representations can learn novel concepts quickly and effectively. This also suggests that, in the future, a more systematic study of the asymptotic properties of supervised training is warranted. In fact, it is possible that certain well defined but broad problems, such as the detection of certain parts in *all* animals, could be solved essentially by "exhaustion", by collecting once for all a sufficiently large pool of annotated examples.

Acknowledgments. We would like to thank Xerox Research Center Europe and ERC 677195-IDIU for supporting this research.

References

[1] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *Proc. ICCV*, pages 2252–2259. IEEE, 2011.

[2] Mahsa Baktashmotlagh, Mehrtash Harandi, Brian Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proc. ICCV*, pages 769–776, 2013.

[3] Yee Seng Chan and Hwee Tou Ng. Domain adaptation with active learning for word sense disambiguation. In *annual meeting - association for computational linguistics*, volume 45, page 49, 2007.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proc. CVPR*, pages 1971–1978, 2014.

[5] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proc. CVPR*, 2015.

[6] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[7] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. Towards scalable dataset construction: An active learning approach. In *Proc. ECCV*, pages 86–98. Springer, 2008.

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

[9] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *Proc. ECCV*, pages 452–466, 2010.

[10] Ian Endres, Vivek Srikumar, Ming-Wei Chang, and Derek Hoiem. Learning shared body plans. In *Proc. CVPR*, pages 3130–3137, 2012.

[11] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proc. CVPR*, 2009.

[12] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[13] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proc. ICCV*, pages 2960–2967, 2013.

[14] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proc. CVPR*, 2016.

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *Proc. ICML*, 2015.

[16] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2551–2559, 2015.

[17] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, pages 2066–2073. IEEE, 2012.

[18] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. ICCV*, pages 999–1006. IEEE, 2011.

[19] M. Guillaumin, D. Küttel, and V. Ferrari. Imagenet auto-annotation with segmentation propagation. *IJCV*, 2014.

[20] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *Proc. ECCV*, 2012.

[21] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. 2013.

[22] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Proc. NIPS*. 2014.

[23] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[24] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *Proc. CVPR*, pages 2372–2379. IEEE, 2009.

[25] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *Proc. ICCV*, pages 1–8. IEEE, 2007.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL http://arxiv.org/abs/1602.07332.

[27] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Proc. NIPS*, pages 231–238. MIT Press, 1995.

[28] D. Küttel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *Proc. CVPR*, 2012.

[29] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, pages 951–958. IEEE, 2009.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. ICCV*, pages 3431–3440, 2015.

[31] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Proc. NIPS*, pages 1601–1609. 2014.

[32] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010. http://www.milbo.org/muct.

[33] Damian Mrowca, Marcus Rohrbach, Judy Hoffman, Ronghang Hu, Kate Saenko, and Trevor Darrell. Spatial semantic regularisation for large scale object detection. In *Proc. ICCV*, 2015.

[34] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proc. ICML*, pages 10–18, 2013.

[35] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *Proc. ECCV*, pages 354–368. Springer, 2012.

[36] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, and Hong-Jiang Zhang. Two-dimensional active learning for image classification. In *Proc. CVPR*, pages 1–8. IEEE, 2008.

[37] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics, 2010.

[38] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where – and why? semantic relatedness for knowledge transfer. In *Proc. CVPR*, 2010.

[39] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *Proc. NIPS*, 2013.

[40] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proc. CVPR*, pages 2121–2131, 2015.

[41] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, pages 213–226. Springer, 2010.

[42] Avishek Saha, Piyush Rai, Hal Daumé III, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[43] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[45] Michael Stark, Michael Goesele, and Bernt Schiele. A shape-based object class model for knowledge transfer. In *Proc. ICCV*, 2009.

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015.

[47] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *Proc. NIPS*, 2011.

[48] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *JMLR*, 2:45–66, 2002.

[49] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *PAMI*, 29(5):854–869, 2007.

[50] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proc. CVPR*, pages 1510–1519. IEEE, 2015.

[51] Shubham Tulsiani, João Carreira, and Jitendra Malik. Pose induction for novel object categories. In *Proc. ICCV*, 2015.

[52] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[53] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proc. ICCV*, pages 4068–4076, 2015.

[54] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *PAMI*, 31(11):2032–2047, November 2009.

[55] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, B. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed. Understanding objects in detail with fine-grained attributes. In *Proc. CVPR*, 2014.

[56] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Proc. CVPR*, pages 2262–2269. IEEE, 2009.

[57] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *Proc. ICCV*, pages 2524–2531. IEEE, 2011.

[58] Xuezhi Wang and Jeff Schneider. Flexible transfer learning under support and model shift. In *Proc. NIPS*, pages 1898–1906, 2014.

[59] Xuezhi Wang, Tzu-Kuo Huang, and Jeff Schneider. Active transfer learning under model shift. In *Proc. ICML*, pages 1305–1313, 2014.

[60] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197. ACM, 2007.

[61] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2013.

[62] Yongxin Yang and Timothy M. Hospedales. Multivariate regression on the grassmannian for predicting novel domains. In *Proc. CVPR*, 2016.

[63] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, pages 94–108, 2014.