15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Hierarchical Zero-Shot Classification with Convolutional Neural Network Features and Semantic Attribute Learning

Jared Markowitz, Aurora C. Schmidt, Philippe M. Burlina, I-Jeng Wang
Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD, USA
Jared.Markowitz@jhuapl.edu

Abstract

We examine hierarchical approaches to image classification problems that include categories for which we have no training examples. Building on prior work in hierarchical classification that optimizes the trade-off between depth in a tree and accuracy of placement, we compare the performance of multiple formulations of the problem on both previously seen (non-novel) and previously unseen (novel) classes. We use a subset of 150 object classes from the ImageNet ILSVRC2012 data set, for which we have 218 human-annotated semantic attribute labels and for which we compute deep convolutional features using the OVERFEAT network. We quantitatively evaluate several approaches, using input posteriors derived from distances to SVM classifier boundaries as well as input posteriors based on semantic attribute estimation. We find that the relative performances of the methods differ in non-novel and novel applications and achieve information gains in novel applications through the incorporation of attribute-based posteriors.

1 Introduction

Real world machine learning applications increasingly address large and diverse data sets. While this leads to an abundance of some classes of data, it also uncovers objects that defy categorization in previously seen classes. The task of classifying samples taken from categories where no training examples exist is known as *zero-shot learning*. This problem has interest both from the practical standpoint of automatically labeling novel items (thereby saving the time needed to retrain classifiers) and from the scientific standpoint of understanding how humans perform the task [1].

In this study, we evaluate and combine two different approaches to zero-shot learning: hierarchical classification and mapping to semantic attributes. Hierarchical classification can be applied to previously unseen classes by trading specificity for accuracy. The idea is that while a novel object cannot be classified exactly, correctly placing it in a more general category can still provide useful information. For example, a classifier that has not been trained on peaches may still recognize a peach as a fruit. In contrast, approaches based on semantic attributes leverage the capability of humans to categorize yet unseen classes by learning mappings from image features to human recognizable attributes. It is assumed that while training data for novel classes do not exist, knowledge of their semantic attributes is available. For instance we may have a set of animal images that does not include tigers but still be aware of this large, striped, feline class. The objective is to learn attribute classifiers that transfer to novel images well enough to produce accurate class predictions.

Hierarchical and attribute-based classification methods provide different types of information; the former gives a projected position in an established hierarchy while the latter provides a ranked listing of potential classes and their estimated probabilities. We find that, when allowing for novel classes, the rank of the true class obtained from an attribute-based classifier is generally smaller (better) than the number of potential classes remaining after hierarchical classification. However the former gives no notion of hierarchical context; in fact the top-ranked classes may come from very different regions of a ground truth hierarchy. Hence the two approaches have complementary strengths.

To leverage these different benefits, we investigate methods for combining the two approaches. We work in both directions. Starting with attributes, we show that the posteriors generated from attribute-based classification can be used as input to hierarchical classifiers in order to achieve an increase in average information gained in zero-shot applications. Conversely, starting with hierarchies, we demonstrate an advantage in using hierarchical classifiers to pare down the ranked lists produced by attribute-based analyses. Throughout we consider only "blind" zero-shot applications, or those where it is not known whether a given test image comes from a previously seen or unseen class.

2 Related Work

Many approaches for hierarchical classification have recently been published [2, 3, 4, 5]. One particularly elegant method for hierarchical classification is the Dual Accuracy Reward Trade-off Search (DARTS), described in [6] and below. Here we utilize DARTS as well as the closely-related Maximum Expected Reward (MAX-EXP) approach, also described in [6].

Recent works on zero-shot classification include [7, 8, 9, 10]. In this study we use class rankings and attribute posterior probabilities similar to the direct and indirect methods in [11].

One previous work incorporating attribute estimation into a hierarchical approach is [12]. In [12], sets of both distinguishing attributes and refined attribute classifiers are specified at each inner node. This enables improved zero-shot classification performance, although in a manner that differs from our approach.

3 Zero-Shot Classification Methods

The hierarchical classification and attribute-based approaches employed in this work are described below.

3.1 Hierarchical Classification

The DARTS algorithm [6] trades off accuracy and specificity to generate appropriate classifications for objects of varying certainties. It finds a classifier that solves the optimization problem

$$\begin{aligned} \underset{f}{\text{maximize}} \quad & R(f) \\ \text{subject to} \quad & \Phi(f) \geqslant 1-\epsilon, \end{aligned} \tag{1}$$

where f is the chosen classifier, $R(f)$ is the average reward, $\Phi(f)$ is the expected likelihood that the true class is located at or is a descendant of the chosen node of the hierarchy, and $1-\epsilon$ is the desired minimum accuracy. Here the reward is defined in terms of information gained by reducing the number of potential bottom level "leaf" classes.

In [6], the problem is parameterized via a Lagrange Multiplier. For classifier f and multiplier λ, the Lagrange function $L(f,\lambda)$ is given by

$$L(f,\lambda) = R(f) + \lambda(\Phi(f) - 1 + \epsilon). \tag{2}$$

The optimal λ is determined through a binary search, as it can be shown that $\Phi(f)$ and $R(f)$ are non-decreasing in opposite directions. The resulting optimal classifier is given by

$$f_\lambda(x) = \underset{v\in\mathcal{V}}{\operatorname{argmax}}(r_v + \lambda)p_{Y|X}(v|x), \tag{3}$$

where r_v is the reward for node v in the hierarchy $\mathcal{V}$ and $p_{Y|X}(v|x)$ is the posterior probability of an image described by x being in or a descendant of node v. This posterior can be obtained in different ways. In [6], one-vs-all linear SVM classifiers are trained on the set of training classes Y, which are the leaf nodes of the hierarchy. Probability distributions are generated by applying Platt Scaling [13] to Y and summing up the tree. Note that the parameter λ is learned using both the training data and the true training classification labels, accounting for some tuning of trust in the posterior estimates.

In addition to DARTS, we investigated a more direct approach to solving (1). This MAX-EXP method (also from [6]) calls for selection of the node with highest expected reward, subject only to a threshold in posterior probability. This threshold is learned to guarantee a pre-defined average accuracy in the validation data. Formally, this classifier is defined by

$$f_\theta(x) = \underset{v\in\mathcal{V}:p_{Y|X}(v|x)>\theta}{\operatorname{argmax}} r_v p_{Y|X}(v|x), \tag{4}$$

where θ is the posterior probability threshold.

Finding optimal values for the parameters λ and θ can be difficult. In the case of λ, the upper bound is determined by the smallest ratio of true class posterior to another posterior in the training data. When erroneous posteriors are present, arbitrarily large λ values will be required to achieve high accuracy. This can result in the classifier often pushing all inputs to the top/root node, resulting in uninformative choices. The same behavior occurs for MAX-EXP, but requires θ to be pushed arbitrarily close to 1 to achieve high required accuracies.

Another notable issue with both DARTS and MAX-EXP is the lack of uniformity in their performance across class. To guarantee an overall accuracy, the algorithms will often neglect more difficult classes while maximizing their performance on easier classes. While we do not tackle it here, this issue may be addressed by adjusting the rewards to emphasize different branches of the tree.

3.2 Classification based on Semantic Attributes

Semantic attributes provide a bridge from automatically generated image features to human intuition. A learned mapping from features to attributes can be applied to both non-novel and novel classes, enabling zero-shot learning. Here we focus on the direct and indirect approaches discussed in [11, 14]. We define the feature space as $\mathcal{X}$, the set of non-novel labels as $\mathcal{Y}$, and the set of novel labels as $\mathcal{Z}$.

In the direct method, a classifier is learned for each attribute from the examples in the training data. We use 3-class classifiers for each attribute, allowing for both yes and no states (denoted 1 and -1), as well as an in-between or "not-applicable" state (0). We start with a set of labeled examples $(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_n, l_n) \in \mathcal{X} \times \mathcal{Y}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are deep convolutional neural network features. We then use the class-to-attribute matrix $\mathbf{V}$ to create a set of attribute-labeled training examples $(\mathbf{x}_1, u_1), \ldots, (\mathbf{x}_n, u_n) \in \mathcal{X} \times \{-1, 0, +1\}$ for each attribute $j = 1, \ldots, N_a$. We build N_a attribute classifier models from these data using Linear Support Vector Machines (LSVM) and use them to infer the feature to attribute mapping $\hat{\mathbf{v}}(\mathbf{x})$.

In the indirect method, we first learn the posterior distribution on class given features, $P(Y|\mathbf{X})$. This is achieved through one-vs-all LSVM classification followed by Platt sigmoid scaling [13]. $P(Y|\mathbf{X})$ is translated to attribute estimates through a weighted average:

$$E_P[v_j] = \sum_{i\in Y} P(\text{class} = i|\mathbf{X} = \mathbf{x})v_{i,j}. \tag{5}$$

The final value for each attribute estimate is then thresholded to $\{-1, 0, +1\}$.

To generate class estimates and probability distributions (both non-novel and novel), we employ the maximum likelihood (ML) method in [11]. This method uses the measured error rates of each attribute classifier in a validation data set as well as assumed independence of errors to compute posterior likelihoods of class given the inferred attribute vector. Given an estimated attribute vector $\mathbf{v}$, the likelihood of class i is given by

$$P(\mathbf{x}|y_i) \sim P(\hat{\mathbf{v}} = \mathbf{v}_i) \approx \prod_{j=1,..,N_a} P(\hat{v}(j)|v_{i,j}), \tag{6}$$

where the $\mathbf{v}_i$ are the attribute vectors in $\mathbf{V}$ corresponding to the class $i \in \mathcal{A} = \mathcal{Y} \cup \mathcal{Z}$. Previous experiments showed generally equivalent or better performance of ML class ranking to distance-based ranking. Critical to our purposes here, the ML method provides class posterior distribution estimates.

3.3 Combination Approaches

These attribute-based classification schemes may be incorporated into hierarchical classification in at least

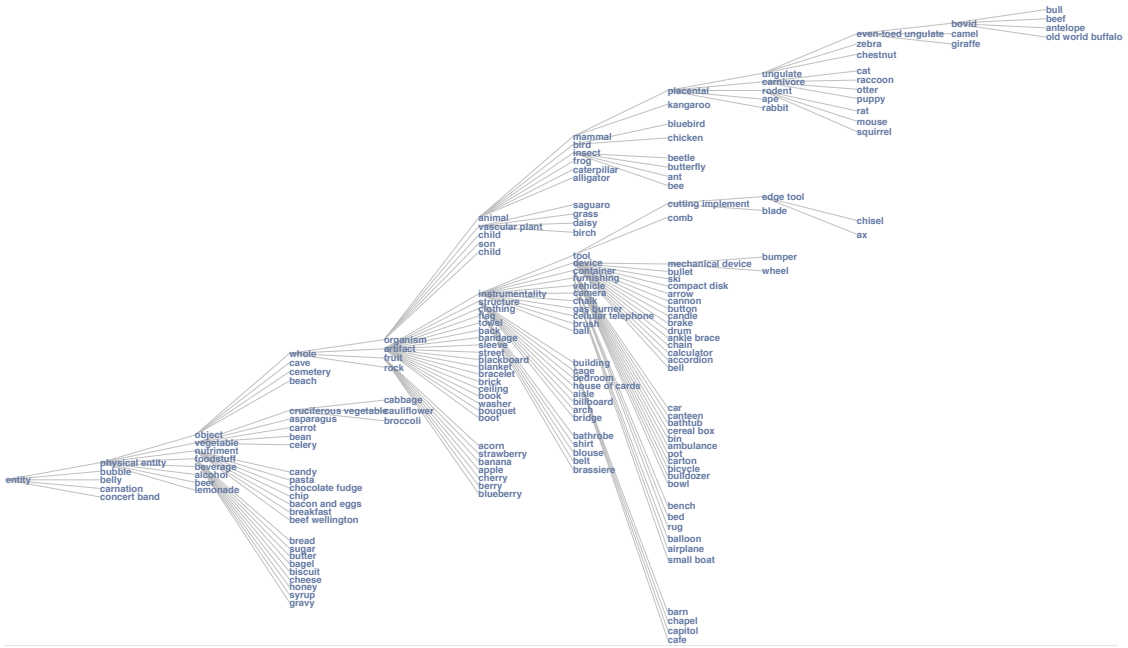

Figure 1. Ground truth hierarchy used in experiments, generated using clustering of leaf node attributes and subsequent cross-referencing with WordNet[15].

two ways. First, the class posterior probability estimates derived from the attribute analysis may be used as the inputs to the hierarchical classifier. The attribute-based posteriors offer additional semantic information to the classifiers, potentially improving zero-shot classification performance. Second, the hierarchical classifier can be used to choose the number N of leaf nodes to consider from the class rankings provided by the attribute-based approach. We refer to this latter approach as "TOPN."

4 Experiments and Results

4.1 Data Processing

The data set chosen for this analysis was comprised of 192,870 images from 150 ImageNet [16] classes. The images were split into training, validation, and testing sets according to a $90\% - 5\% - 5\%$ split. Of the original 150 classes, 30 were held out of training for zero-shot analysis. All images were run through the OVERFEAT deep convolutional network [17], producing a set of 4096-dimensional feature vectors.

Each class was additionally endowed with a set of 218 human-recognizable attributes [7]. The ground truth hierarchies required for the hierarchical classifiers were generated through clustering of these attribute vectors. Each data class was taken to be a bottom level node of the hierarchy. The pairwise correlation distances of the semantic attribute vectors for each class were computed and clustered to identify higher level groupings [18]. Each higher level node was matched to the WordNet[15] node that represented the nearest common parent of all its children (Figure 1).

The posterior class probability estimates required for conventional hierarchical classification were generated as follows. First the feature vectors were used to train one-vs-all linear SVM classifiers for each training class using the LIBLINEAR default C=1 [19] (instead of the C=100 values used in [6]). Platt scaling [13] was then applied to generate leaf node probability estimates for each input image. The probabilities of higher level nodes were estimated through summation of the probabilities of their children, moving up the hierarchy. For this summation the tree was pruned to remove the 30 novel classes, consistent with [6].

4.2 Comparison of Hierarchical Approaches

These data were used to evaluate the hierarchical classification strategies described in Section 3. The effects of using posterior probabilities generated by the attribute analysis (6) in place of the class-based posterior probabilities as classifier inputs as well as the "TOPN" cross-referencing approach were quantified. When attribute-based posteriors were used the full hierarchy (including novel classes) was employed, in contrast to conventional hierarchical classification.

To characterize the different methods, we plotted the average reward versus the average containment accuracy observed in the testing set. As in [6], reward for an individual image was defined in terms of information gain:

$$r(v) = \log_2 |\mathcal{Y}| - \log_2 \sum_{y \in \mathcal{Y}} [v \in \pi(y)] \,. \qquad (7)$$

Again v represents the chosen node and $\mathcal{Y}$ the set of leaf nodes. The set of ancestors of a leaf node y is $\pi(y)$. Traces were generated by considering a range of hierarchical classifier parameters (λ for DARTS, θ for MAX-EXP).

The reward-accuracy curves for the various methods applied to previously seen and unseen classes are shown in Figures 2 and 3, respectively. As expected, performance is far superior in classes for which there is training data. Nonetheless, as in [6], the hierarchical classifiers are able to provide a non-negligible amount of information when applied to novel classes. Notably the relative performance of different classifiers differs in non-novel and novel applications. Many methods provide comparable performance for non-novel classes, with standard DARTS narrowly offering the highest average reward for high containment accuracy. In the

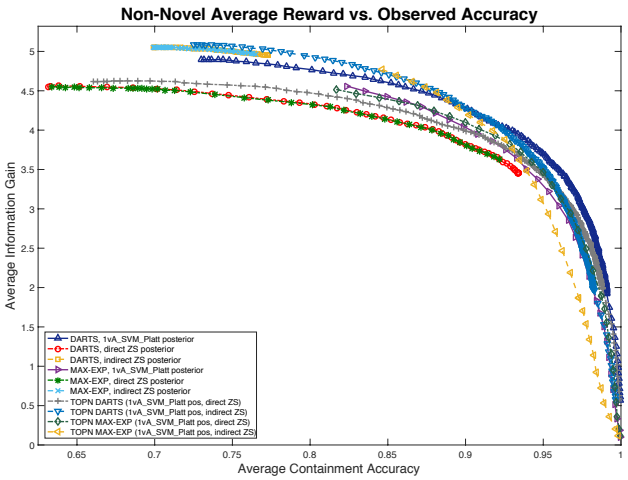

Figure 2. Hierarchical classification of images from previously seen (non-novel) classes.

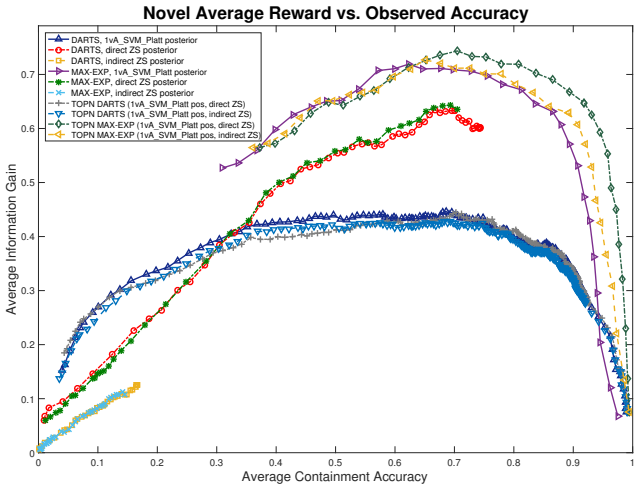

Figure 3. Hierarchical classification of images from previously unseen (novel) classes.

novel case, MAX-EXP generally outperforms DARTS. The TOPN MAX-EXP method shows the benefit of incorporating rankings from both the direct and indirect approaches, with the rankings from the direct approach being more advantageous. The performance of DARTS on novel classes is seen to be significantly improved[1] for moderate accuracy levels through the use of posteriors derived from direct zero-shot methods as compared to conventional class-based posteriors.

5 Conclusion

We have evaluated and combined multiple methods for hierarchical and attribute-based classification, quantifying performance on previously seen and unseen classes separately. We see different relative performances of methods in the two cases. While most approaches perform similarly on previously seen classes, significant information gains on previously unseen classes are achieved by augmenting conventional hierarchical classifiers with attribute-based input.

[1]The form of the reward metric leads to differences in small rewards representing much larger reductions in the number of potential leaf classes than the same differences in high rewards.

Acknowledgment

We thank the authors of [7] for providing the class to attribute mapping of their data set.

References

[1] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[2] O. Dekel, J. Keshet, and Y. Singer, "Large margin hierarchical classification," in *ICML*, 2004.

[3] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Incremental algorithms for hierarchical classification," *Journal of Machine Learning Research*, vol. 7, pp. 31–54, 2006.

[4] B. Zhao, L. Fei-Fei, and E. P. Xing, "Large-scale category structure aware image categorization," in *NIPS*, 2011.

[5] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv:1612.08242v1*, 2016.

[6] J. Deng, J. Krause, A. Berg, and L. Fei-Fei, "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition," in *CVPR*, 2012.

[7] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *NIPS*, 2009, pp. 1410–1418.

[8] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," in *NIPS*, 2013.

[9] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *PAMI*, vol. 36, no. 3, pp. 453–465, 2014.

[10] S. Antol, Z. C. Lawrence, and P. Devi, "Zero-shot learning via visual abstraction," in *ECCV*, 2014.

[11] P. M. Burlina, A. C. Schmidt, and I.-J. Wang, "Zero shot deep learning from semantic attributes," in *ICMLA*, 2015, pp. 871–876.

[12] Z. Al-Halah and R. Stiefelhagen, "How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes," in *WACV*, Jan 2015, pp. 837–843.

[13] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[14] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*. IEEE, 2009, pp. 951–958.

[15] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv:1312.6229*, 2013.

[18] D. Mullner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv:1109.2378v1*, 2011.

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.