Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition

Jia Deng[1,2], Jonathan Krause[1], Alexander C. Berg[3], Li Fei-Fei[1]
Stanford University[1], Princeton University[2], Stony Brook University[3]

Abstract

As visual recognition scales up to ever larger numbers of categories, maintaining high accuracy is increasingly difficult. In this work, we study the problem of optimizing accuracy-specificity trade-offs in large scale recognition, motivated by the observation that object categories form a semantic hierarchy consisting of many levels of abstraction. A classifier can select the appropriate level, trading off specificity for accuracy in case of uncertainty. By optimizing this trade-off, we obtain classifiers that try to be as specific as possible while guaranteeing an arbitrarily high accuracy. We formulate the problem as maximizing information gain while ensuring a fixed, arbitrarily small error rate with a semantic hierarchy. We propose the Dual Accuracy Reward Trade-off Search (DARTS) algorithm and prove that, under practical conditions, it converges to an optimal solution. Experiments demonstrate the effectiveness of our algorithm on datasets ranging from 65 to over 10,000 categories.

1. Introduction

Even conservative estimates suggest that there are tens of thousands of object classes in our visual world [3]. This number may easily scale up by orders of magnitude considering more fine-grained classes. Can computers recognize all object classes while almost never making mistakes, a challenging task even to a knowledgeable human?

This seems elusive, given that the state of the art performance on 10K-way classification is only 16.7% [26]. Of course there *is* a way to *always be right*; we can just report everything as an "entity", which is, however, not very informative. This paper shows how to achieve something sensible between the two extremes of inaccurate choices forced among a large number of categories and the uninformative option of declaring that everything is an "entity".

One key to success is to observe that object categories form a semantic hierarchy, consisting of many levels of abstraction. For example, a kangaroo is also a mammal, an animal, and a living thing. The classifier should predict "mammal" instead if it is uncertain of the specific species. Meanwhile, the classifier should try to be as specific as possible. Consider the bottom image in Fig. 1, where it would have been correct to report animal, but choosing mammal provides more information without being wrong. A sensible classifier thus "hedges its bets" as necessary, maintaining high accuracy while making its best effort for specificity.

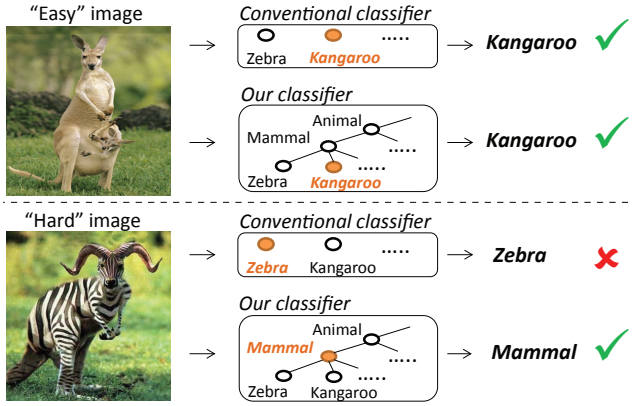

Figure 1. Conventional classifier versus our approach.

Our goal is *to create a classification system that maximizes information gain while maintaining a fixed, arbitrarily small error rate.* We measure information gain in the standard information theoretical sense, *i.e.*, the decrease in uncertainty from our prior distribution to the posterior over the classes. For example, our prior can be uniform among the tens of thousands of leaf nodes in a hierarchy. A classification output of "mammal", though maintaining uncertainty about the specific species, provides information by ruling out many other possibilities. Note that our algorithmic approach can also handle alternate, application-specific measures instead of information gain.

Results on datasets ranging from 65 to over 10,000 classes show that not only is our proposed algorithm effective at training classifiers that optimize information gain while maintaining high accuracy, but that the resulting classifications are informative. This is a step toward more widely useful classification by making explicit the *trade-off between accuracy and specificity*. This trade-off can be rel-

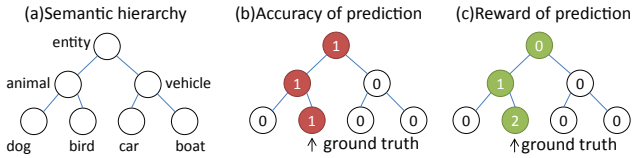

Figure 2. Illustration of the formulation with a simple hierarchy. The numbers correspond to the accuracy (middle) and the reward (information gain) of a prediction (right) given the ground truth.

evant in many visual tasks with high level semantic labels, *e.g.*, detection, scene understanding, describing images with sentences, etc. Our focus here, multiclass image classification, serves as a building block. To our knowledge this is the first time that *optimizing* the accuracy-specificity trade-off has been addressed in large scale visual recognition.

In this work, we make the following contributions: (1) introducing the problem of classification in a hierarchy subject to an accuracy bound while maximizing information gain (or other measures), (2) proposing the Dual Accuracy Reward Trade-off Search (DARTS) algorithm and proving a non-trivial result that, under practical conditions, it converges to an optimal solution, and (3) validating our algorithm with experiments on 65 to more than 10,000 classes, showing large improvements over baseline approaches.

2. Related Work

Our problem is related to cost-sensitive classification and hierarchical classification [33, 22, 7, 9, 4, 2, 14, 1, 35, 29, 20, 21, 16, 23, 34, 5, 18]. The key differences are: (1) conventional multiclass or cost-sensitive techniques do not consider overlapping classes in a hierarchy; (2) previous work on hierarchical classification has not addressed the issue of automatically selecting the appropriate level of abstraction to *optimize* the accuracy-specificity trade-off.

Also related is classification with reject options, which grants binary classifiers an option to abstain, for a particular cost [32, 10]. In the multiclass case [17, 15, 6], also termed "class selective rejection", the classifier can output an *arbitrary* set of classes. Our problem fits in this framework, with the admissible sets restricted to internal nodes of the hierarchy. To our knowledge we are the first to connect class selective rejection with hierarchical visual classification. Our primal dual framework follows [15], but our results on the optimality of DARTS are new (Sec. 4.2).

Our work is also inspired by an emerging trend in computer vision studying large scale visual recognition [19, 27, 13, 12, 31, 7, 26, 24]. Our technique scales up easily and we demonstrate its effectiveness on large scale datasets.

3. Formulation

We describe the visual world with a semantic hierarchy $H = (V, E)$, a directed acyclic graph (DAG) with a unique root $\hat{v} \in V$, each node $v \in V$ representing a semantic class (Fig. 2a). The leaf nodes $\mathcal{Y} \subset V$ are *mutually exclusive* classes. The internal nodes are unions of leaf nodes determined by the hierarchy, *e.g.*, in Fig. 2a, "animal" is a combination of "dog" and "bird", while "entity" is a combination of everything under "animal" and "vehicle".

Given the hierarchy, it is then correct to label an image at either its ground truth leaf node or any of its ancestors (Fig. 2b), *e.g.*, a dog is also an animal and an entity. Let X be an image represented in some feature space and Y its ground truth leaf label, X and Y drawn from a joint distribution on $\mathcal{X} \times \mathcal{Y}$. A classifier $f : \mathcal{X} \to V$ labels an image $x \in \mathcal{X}$ as a node $v \in V$, either a leaf node or an internal node. The accuracy $\Phi(f)$ of the classifier f is then

$$\Phi(f) = \mathbb{E}\left[(f(X) \in \pi(Y)\right]^{1}, \tag{1}$$

where $\pi(Y)$ is the set of all possible correct predictions, *i.e.*, the ground truth leaf node and its ancestors. Note that without the internal nodes, $\Phi(f)$ reduces to the conventional flat multiclass accuracy. In this paper, we use "accuracy" in the hierarchical sense unless stated otherwise.

The conventional goal of classification is maximizing accuracy. In our case, however, always predicting the root node ensures 100% accuracy, yielding an uninformative solution. We clearly prefer an answer of "dog" over "entity", whenever they are both correct. We encode this preference as a reward $r_v \geq 0$ for each node $v \in V$. One natural reward is information gain, the decrease in uncertainty (entropy) from the prior distribution to the posterior over the leaf classes. Assuming a uniform prior, it is easy to verify that a prediction at node v decreases the entropy by

$$r_v = \log_2 |\mathcal{Y}| - \log_2 \sum_{y \in \mathcal{Y}} [v \in \pi(y)]. \tag{2}$$

The information gain is zero at the root node and maximized at a leaf node. Note that we use information gain in experiments but our algorithm and analysis can accommodate an *arbitrary* non-negative reward. Given the reward of each node, the reward $R(f)$ for a classifier f is

$$R(f) = \mathbb{E}\left(r_{f(X)}[f(X) \in \pi(Y)]\right), \tag{3}$$

i.e., r_v for a correct prediction at node v, and 0 for a wrong one (Fig. 2c). In the case of information gain, the reward of a classifier is the average amount of correct information it gives. Our goal then is to maximize the reward given an *arbitrary* accuracy guarantee $0 < 1 - \epsilon \leq 1$, *i.e.*,

$$\begin{aligned} \underset{f}{\text{maximize}} \quad & R(f) \\ \text{subject to} \quad & \Phi(f) \geq 1 - \epsilon. \end{aligned} \tag{OP1}$$

Note that OP1 is always feasible because there exists a trivial solution that only predicts the root node.

[1] "$[P]$" is the Iverson bracket, *i.e.*, 1 if P is true and 0 otherwise.

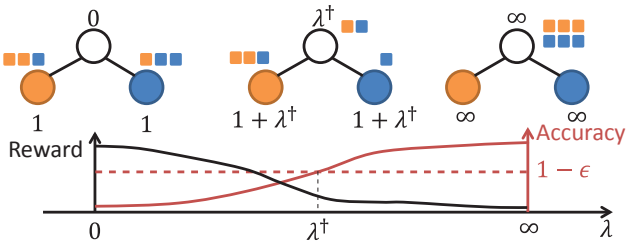

Figure 3. **Bottom:** The general properties of the reward and accuracy of f_λ, a classifier that maximizes the Lagrange function, with respect to the dual variable λ. An optimal solution to OP1 is $f_{\lambda^\dagger}$, where the accuracy is exactly the minimum guaranteed, provided $\lambda^\dagger$ exists. **Top:** The solid squares represent image examples; their color indicates the ground truth. The numbers next to the nodes are the transformed rewards $r_v + \lambda$ in the Lagrange function. As λ increases, the classifier f_λ predicts more examples to the root node. Eventually every example goes to the root node unless some other node already has posterior probability 1.

4. The DARTS Algorithm

In this section we present the Dual Accuracy Reward Trade-off Search (DARTS) algorithm to solve OP1 and prove its optimality under practical conditions.

DARTS is a primal dual algorithm based on "the generalized Lagrange multiplier method" [11]. In our case, the dual variable controls the trade-off between reward and accuracy. We first write the Lagrange function

$$L(f, \lambda) = R(f) + \lambda(\Phi(f) - 1 + \epsilon), \tag{4}$$

with the dual variable $\lambda \geq 0$. Given a λ, we obtain a classifier f_λ that maximizes the Lagrange function, a weighted sum of reward and accuracy controlled by λ. It can be shown that the accuracy of the classifier $\Phi(f_\lambda)$ is non-decreasing and the reward $R(f_\lambda)$ non-increasing with respect to λ [11]. Moreover, if a $\lambda^\dagger \geq 0$ exists such that $\Phi(f_{\lambda^\dagger}) = 1 - \epsilon$, *i.e.*, the classifier $f_{\lambda^\dagger}$ has an accuracy of exactly $1-\epsilon$, then $f_{\lambda^\dagger}$ is optimal for OP1 [11]. These properties, illustrated in Fig. 3 (bottom), lead to a binary search algorithm to find such a $\lambda^\dagger$. At each step the algorithm seeks a classifier that maximizes the Lagrange function. It converges to an optimal solution provided such a $\lambda^\dagger$ exists.

To apply this framework, however, we must address two challenges: (1) finding the classifier that maximizes the Lagrange function and (2) establishing conditions under which $\lambda^\dagger$ exists and thus the binary search converges to an optimal solution. The latter is particularly non-trivial as counterexamples exist, *e.g.*, the red curve in Fig. 3 can be discontinuous and as a result the dashed line can fail to meet it.

4.1. Maximizing the Lagrange Function

DARTS maximizes the Lagrange function by using posterior probabilities. Using Eqn. 3, Eqn. 1, and Eqn. 4 yields

$$L(f, \lambda) = \mathbb{E}\ (r_{f(X)} + \lambda)[f(X) \in \pi(Y)] + \lambda(\epsilon - 1), \tag{5}$$

i.e., maximizing the Lagrange function is simply maximizing a transformed reward $r_v + \lambda, \forall v \in V$. This can be achieved by estimating posterior probabilities and predicting the node with the maximum expected reward, breaking ties arbitrarily. Let f_λ be such a classifier given a λ, then

$$f_\lambda(x) = \operatorname*{argmax}_{v \in V} (r_v + \lambda) p_{Y|X}(v|x), \tag{6}$$

where $p_{Y|X}(v|x) = \Pr(v \in \pi(Y)|X = x)$. This can be easily proven by rewriting Eqn. 5 using iterated expectations, conditioning on X first.

Let's examine f_λ. When $\lambda = 0$, f_0 simply maximizes the original reward. As λ tends to infinity, the transformed reward $r_v + \infty$ becomes equal on all nodes. The root node has maximum probability and therefore the best expected reward. Thus every example is predicted to the root node, unless some other node already has probability 1. Either way, all predictions are accurate with $\lambda = \infty$.

To obtain the posterior probabilities, we learn conventional one-vs-all classifiers on the leaf nodes (*e.g.*, SVMs), obtain probability estimates (*e.g.*, via Platt scaling [25]), and sum them to get internal node probabilities.

We summarize DARTS in Algorithm 1. It first obtains posterior probabilities for all nodes and exits if f_0, the classifier that maximizes the original reward only, is already at least $1 - \epsilon$ accurate (step 1–4). Otherwise it does a binary search to find a $\lambda^\dagger > 0$ such that the classifier that maximizes the transformed ward $r_v + \lambda^\dagger$ is exactly $1 - \epsilon$ accurate. The upper bound of the binary search interval, $\bar{\lambda}$, is set such that $\lambda^\dagger \leq \bar{\lambda}$ is guaranteed (proof in the supplemental material). DARTS runs for no more than T iterations or until $\Phi(f_\lambda)$ is within a small number $\tilde{\epsilon}$ from $1 - \epsilon$.

Algorithm 1 DARTS

1. Obtain $p_{Y|X}(y|x), y \in \mathcal{Y}$.
2. $p_{Y|X}(v|x) \leftarrow \sum_{y \in \mathcal{Y}} [v \in \pi(y)] p_{Y|X}(y|x), \forall v \in V$.
3. $f_0 \leftarrow \arg\max_{v \in V} r_v p_{Y|X}(v|x)$.
4. If $\Phi(f_0) \geq 1 - \epsilon$, return f_0.
5. $\bar{\lambda} \leftarrow (r_{max}(1-\epsilon) - r_{\hat{v}})/\epsilon$, where $r_{max} = \max_{v \in V} r_v$.
6. Do binary search for a $\lambda \in (0, \bar{\lambda}]$ until $0 \leq \Phi(f_\lambda) - 1 + \epsilon \leq \tilde{\epsilon}$ for a maximum of T iterations. Return f_λ.

To obtain the classifier f_λ given a new λ (step 6), it suffices to have the posterior probabilities on the leaf nodes. Thus we only need to learn 1-vs-all classifiers on the leaf nodes *once*, *i.e.*, DARTS essentially converts a "base" flat classifier with probability estimates to a hierarchical one with the optimal accuracy-specificity trade-off.

Finally we remark that DARTS is not sensitive to non-exact maximization of the Lagrange function, *e.g.*, inaccurate probability estimates, as the error will not be amplified [11]: if a solution f_λ is within $\delta > 0$ from the maximiz-

ing the Lagrange function, then with the accuracy guarantee set to that of f_λ, f_λ is within δ from maximizing the reward.

4.2. Optimality of DARTS

Now we prove that under practical conditions, roughly when the posterior probabilities are continuously distributed, DARTS converges to an optimal solution.

The key is to investigate when the dual variable $\lambda^\dagger$ exists, *i.e.*, when the monotonic red curve in Fig. 3 can meet the dashed line. This is only of concern when $\Phi(f_0) < 1 - \epsilon$, *i.e.*, the start of the red curve is below the dashed line, because otherwise we have satisfied the accuracy guarantee already. With $\Phi(f_0) < 1-\epsilon$, $\lambda^\dagger$ may not exist in two cases: (1) when the end of the curve is below the dashed line, *i.e.*, $\Phi(f_\infty) < 1-\epsilon$, or (2) when the curve is discontinuous. Our main theoretical results state that under normal conditions, these two cases cannot happen and then $\lambda_\dagger$ must exist.

Case(1) cannot happen because we can show that $\bar{\lambda} > 0$ and $\Phi(f_{\bar{\lambda}}) \geq 1 - \epsilon$, where $\bar{\lambda}$ is defined in line 5 of DARTS (proof in the supplemental material).

Case(2) is harder as the curve can indeed be discontinuous (see the supplemental material for a concrete case). However, we can show that case(2) cannot occur if the posterior probabilities are continuously distributed except possibly at 0 or 1, a condition normally satisfied in practice. Consider, for example, a hierarchy of two leaf nodes a and b. The posterior probability $p_{Y|X}(a|X)$, as a function of X, is also a random variable. The condition implies that the distribution of $p_{Y|X}(a|X)$ does not concentrate on any single real number other than 0 and 1, *i.e.*, practically, the posterior probability estimates are sufficiently diverse.

Formally, let $\Delta = \{q \in \mathbb{R}^{|\mathcal{Y}|-1} : q \succeq 0, \|q\|_1 \leq 1\}$ be the set of possible posterior probabilities over the $|\mathcal{Y}|-1$ leaf nodes. Note that for $|\mathcal{Y}|$ leaf nodes there are only $|\mathcal{Y}|-1$ degrees of freedom. Let $\Delta^\ddagger = \{q \in \Delta : \|q\|_\infty = 1 \vee q = 0\}$ be the set of posterior probabilities at the vertices of Δ, where one of the leaf nodes takes probability 1. Let $\vec{p}_{Y|X} : \mathcal{X} \to \Delta$ be a Borel measurable function that maps an example x to its posterior probabilities on leaf nodes. Let $Q = \vec{p}_{Y|X}(X)$ be the posterior probabilities on leaf nodes for the random variable X. As a function of X, Q is also a random variable. Our main result is the following theorem.

Theorem 4.1. *If* $\Pr(Q \in \Delta^\ddagger) = 1$*, or* Q *has a probability density function with respect to the Lebesgue measure on* $\mathbb{R}^{|\mathcal{Y}|-1}$ *conditioned on* $Q \notin \Delta^\ddagger$*, then, for any* $0 \leq \epsilon \leq 1$*, DARTS converges to an optimal solution of OP1.*

Sketch of Proof. We outline the key steps here with the full proof in the supplemental material. The goal is to show the continuity of $\Phi(f_\lambda)$ with respect to λ. We first prove that

$$\Phi(f_\lambda) = p^\ddagger + (1 - p^\ddagger) \sum_{v \in V} \int_{\Gamma_v(\lambda)} q_v p_Q(q) \, \mathrm{d}q,$$

Dataset	Tr	Val	Ts	# Leaf	# Int	H
ILSVRC65	100	50	150	57	8	3
ILSVRC1K	1261	50	150	1000	676	17
ImageNet10K	428	214	213	7404	3043	19

Table 1. Dataset statistics: average number of images per class for training (Tr), validation (Val) and test (Ts), number of leaf and internal nodes, and height of the hierarchy (H).

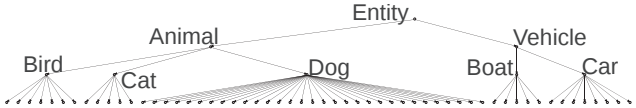

Figure 4. The tree structure of ILSVRC65.

where $p^\ddagger = \Pr(Q \in \Delta^\ddagger)$, $p_Q(q)$ is the (conditional) density function when $Q \notin \Delta^\ddagger$, and $\Gamma_v(\lambda) = \{q \in \Delta : (r_v + \lambda)q_v > (r_{v'} + \lambda)q_{v'}, \forall v' \neq v\}$ is the polyhedron in Δ that leads to a prediction v. We then show the continuity of $\int_{\Gamma_v(\lambda)} q_v p_Q(q) \, \mathrm{d}q$ with respect to λ using Lebesgue's dominated convergence theorem. □

Note that our condition differs from the one given in [15] for strong duality in a general class selective rejection framework, *i.e.*, a continuous density function $p_{X|Y}(x|y) = \Pr(X = x|Y = y)$ exists for each $y \in \mathcal{Y}$. First, neither condition implies the other. Second, theirs guarantees strong duality but not the optimality of a dual algorithm using only posterior probabilities to maximize the Lagrange function, as the maximizer may not be unique. We elaborate using a concrete example in the supplemental material.

In practice, one can estimate whether the condition holds by checking the classifier f_λ DARTS returns. If $\lambda = 0$ or the accuracy of f_λ is close to $1 - \epsilon$, the solution is near optimal. Otherwise $\lambda > 0$ and the accuracy of f_λ is $1-\epsilon' \neq 1 - \epsilon$, in which case the classifier f_λ is sub-optimal for the $1 - \epsilon$ accuracy guarantee, but it is nonetheless optimal for a guarantee of $1 - \epsilon'$ [11].

5. Experiments

Datasets. We use three datasets ranging from 65 to over 10,000 classes: ILSVRC65, ILSVRC1K [28], and ImageNet10K [7]. They are all subsets of ImageNet [8], a database of many classes organized by the WordNet hierarchy. Table 1 lists the statistics. We follow the train/val/test split in [28] for ILSRVC1K. For ImageNet10K, we use a 50-25-25 train/val/test split and exclude images from the internal nodes, as we require that all images have ground truth at leaf nodes. ILSVRC65 is a subset of ILSVRC1K consisting of the leaf nodes of 5 "basic" categories ("dog", "cat", etc. in Fig. 4), with a simplified hierarchy and a down-sampled training set. The smaller scale allows comparison with more baselines and a thorough exploration of parameter space.

Implementation. We represent all images using the LLC [30] features from densely sampled SIFT over a 16K

codebook (10K for ILSVRC65) and a 2 level spatial pyramid (1×1 and 3×3). We train one-vs-all linear SVMs, convert the outputs of each SVM to probabilities via Platt scaling [25], and then L1 normalize them to get multiclass posterior probability estimates [33].

In implementing DARTS, we obtain f_λ using the training set but estimate $\Phi(f_\lambda)$, the expected accuracy of f_λ, using the validation set (step 6). This reduces overfitting. To ensure with high confidence that the true expected accuracy satisfies the guarantee, we compute the .95 confidence interval of the estimated $\Phi(f_\lambda)$ and stop the binary search when the lower bound is close enough to $1 - \epsilon$.

We also implement TREE-DARTS, a variant of DARTS that obtains posterior probabilities differently. It learns one-vs-all classifiers for each internal node to estimate the *conditional* posterior probabilities of the child nodes. It obtains the posterior probability of a node by multiplying all conditional posterior probabilities on its path from the root.

We compare DARTS with five baselines, LEAF-GT, TREE-GT, MAX-REW, MAX-EXP, MAX-CONF.

LEAF-GT is a naive extension of binary classification with a reject option. It takes the posterior probabilities on leaf nodes and predicts the most likely leaf node, if the largest probability is not below a fixed global threshold. Otherwise it predicts the root node. LEAF-GT becomes a flat classifier with threshold 0 and the trivial classifier that only predicts the root node with any threshold above 1.

TREE-GT takes the same *conditional* posterior probabilities in TREE-DARTS but moves an example from the root to a leaf, at each step following the branch with the highest conditional posterior probability. It stays at an internal node if the highest probability is below a fixed global threshold. This represents the decision tree model used in [29].

MAX-REW predicts the node with the best reward among those with probabilities greater than or equal to a threshold. Intuitively, it predicts the most specific node among the confident ones. MAX-EXP is similar to MAX-REW, except that it predicts the node with the best *expected* reward, *i.e.*, its posterior probability times its reward.

MAX-CONF learns a binary, one-vs-all classifier for each node, including all internal nodes except the root node. Given a test image, it predicts the node with the most confident classifier. Despite being intuitive, this baseline is fundamentally flawed. First, assuming accurate confidences, the confidence of a node should never be more than that of its parent, *i.e.*, we can never be more confident that something is a dog than that it is an animal. Thus in theory only the immediate children of the root node get predicted. Second, it is unclear how to satisfy an arbitrary accuracy guarantee—given the classifiers, the accuracy is fixed.

For all threshold-based baselines, a higher threshold leads to higher accuracy and typically less reward in our experiments. Thus, to satisfy a particular accuracy guarantee, we find the best threshold by binary search.

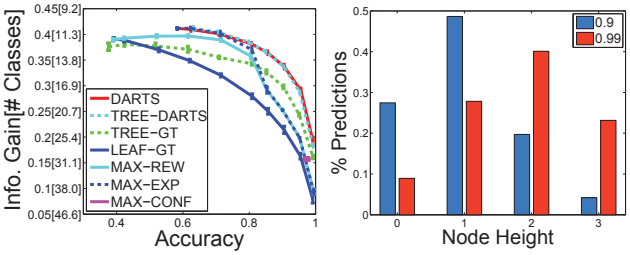

Figure 5. ILSVRC65 results. **Left:** Reward (normalized information gain, with 1 as the maximum possible) versus accuracy. The numbers in brackets on the Y axis indicate the equivalent of number of uncertain classes. The error bars are the standard deviation from 5 training sets, each with 100 images per class randomly sampled from a set of about 1,500 per class. **Right:** The distribution of predictions of DARTS with .9 and .99 accuracy guarantees.

We test all approaches on ILVRC65 but exclude TREE-DARTS, TREE-GT, MAX-CONF on ILSVRC1K and ImageNet10K, because both TREE-DARTS and TREE-GT require significant extension with a non-tree DAG—the child nodes overlap and there can be multiple paths from the root, possibly creating inconsistent probabilities—and because MAX-CONF is fundamentally unusable. We use information gain as reward and normalize it by the maximum possible (*i.e.*, that of leaf nodes) such that the information gain of a flat classifier equals its accuracy.

Results on ILSVRC65. Fig. 5 presents the reward-vs-accuracy curves. We set the accuracy guarantee $1 - \epsilon$ to $\{0, .1, .2, \ldots, .8, .85, .9, .95, .99\}$ and plot the reward and the *actual* accuracy achieved on the test set. Note that all methods are able to satisfy an arbitrary accuracy guarantee, except MAX-CONF that has a fixed accuracy.

First observe that the LEAF-GT curve starts with an accuracy and information gain both at .391, where the global threshold is too low to reject any example, making LEAF-GT equivalent to a flat classifier. The normalized information gain here equals the flat accuracy. In contrast, the DARTS curve starts with an accuracy of .583, achieved by maximizing the reward with a low, inactive accuracy guarantee. This is much higher than the flat accuracy .391 because the rewards on internal nodes already attract some uncertain examples that would otherwise be predicted to leaf nodes. Moreover, DARTS gives more correct information than the flat classifier (.412 versus .391); at this point our classifier is better than a flat classifier in terms of *both* accuracy and information gain. As we increase the accuracy guarantee, specificity is traded off for better accuracy and the information gain drops.

To interpret the information gain, we provide the equivalent number of uncertain leaf classes in Fig. 5 (left). For example, at .9 accuracy, on average DARTS gives the same amount of correct information as a classifier that always correctly predicts an internal node with 14.57 leaf nodes.

Fig. 5 (left) shows that both versions of DARTS significantly beat the baselines, validating our analysis on the optimality of DARTS. Interestingly both versions perform equally well, suggesting that DARTS is not sensitive to the particular means of estimating posterior probabilities.

Fig. 5 (right) plots the distribution of predictions over different semantic levels for DARTS. As the accuracy guarantee increases, the distribution shifts toward the root node. At .9 accuracy, our classifier predicts leaf nodes 27% of the time and one of the 5 basic classes 49% of the time. Given that the flat accuracy is only .391, this is a useful trade-off with a high accuracy and a good amount of information.

Results on ILSVRC1K and ImageNet10K. Fig. 6a and Fig. 6b present the reward-vs-accuracy curves for ILSVRC1K and ImageNet10K. On both datasets, DARTS achieves large improvements over the baselines. Also, at the start of the DARTS curve on ILSVRC1K (*i.e.*, with an inactive accuracy guarantee), DARTS beats the flat classifier (the start of the LEAF-GT curve) on both information gain (.423 versus .415) *and* accuracy (.705 versus .415).

Fig. 6c and Fig. 6d show how the distribution of predictions changes with accuracy for DARTS. As accuracy increases, more examples are predicted to non-root internal nodes instead of leaf nodes. Eventually almost all examples move to the root node. On ILSVRC1K at .9 accuracy, 28% of the examples are predicted to leaf nodes, 55% to non-root internal nodes, and only 17% to the root node (*i.e.*, the classifier declares "entity"). On ImageNet10K, the corresponding numbers are 19%, 64%, and 17%. Given the difficulty of problem, this is encouraging.

Fig. 8 visually compares the confusion matrices of a flat classifier and our classifier, showing that our classifier significantly reduces the confusion among leaf nodes. Fig. 9 shows examples of "hard" images for which the flat classifier makes mistakes whereas our classifier remains accurate.

We remark that in all of our experiments, DARTS either returns $\lambda = 0$ or is able to get sufficiently close to the accuracy guarantee in the binary search, as shown by all trade-off curves. This validates our analysis that, under practical conditions, DARTS converges to an optimal solution.

Zero-shot Recognition. Another advantage of our classifier over a flat one is the ability of zero-shot recognition: classifying images from an unseen class *whose name is also unknown*. The flat classifier completely fails with 0 accuracy and 0 information gain. Our classifier, however, can predict internal nodes to "hedge its bets". Fig. 10 shows the performance of our classifier on 5 randomly chosen classes of ILSVRC65, taken out of the training set *and* the hierarchy. Our classifier is able to predict the correct internal nodes a significant amount of the time and with non-trivial information gain. Our final experiment is recognizing "unusual objects", objects that defy categorization at the subordinate levels. Fig. 7 compares the predictions of a flat classifier versus our classifier, both trained on ILSVRC1K. We observe that the flat classifier is confused whereas our classifier stays sensible.

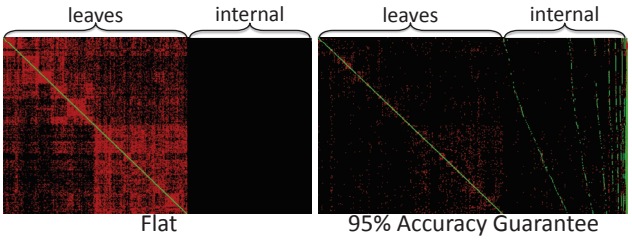

Figure 8. Comparison of confusion matrices on ILSVRC1K classes between a flat classifier and our classifier with a .95 accuracy guarantee. The rows represent leaf nodes; the columns are ordered from leaf to root by node height and then by the DFS order of the hierarchy. The matrices are downsampled; each pixel represents the max confusion between 4×4 entries. Correct predictions are colored green and incorrect ones red.

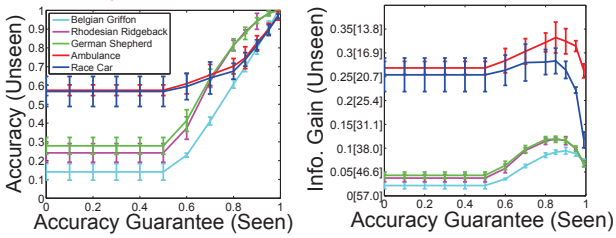

Figure 10. Zero-shot recognition: accuracy and information gain of 5 randomly chosen unseen classes versus accuracy guarantees on seen classes on ILSVRC65.

6. Conclusion

We have introduced the problem of optimizing accuracy-specificity trade-offs in large scale recognition. We have presented the DARTS algorithm, analyzed its optimality, and demonstrated its effectiveness on large scale datasets. This is an encouraging step toward a highly accurate and informative large scale recognition system.

Acknowledgments. We thank Hao Su, Kevin Tang, and Bangpeng Yao for their helpful comments. L.F.-F. is partially supported by NSF CAREER grant (IIS-0845230), the DARPA CSSG grant, and a Google research award. A.C.B. is partially supported by the Stony Brook University Office of the Vice President for Research.

References

[1] Y. Amit, M. Fink, and N. Srebro. Uncovering shared structures in multiclass classification. In *ICML*, 2007.

[2] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2010.

[3] I. Biederman. Recognition by components: A theory of human image understanding. *PsychR*, 94(2):115–147, 1987.

[4] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Incremental algorithms for hierarchical classification. *JMLR*, 7:31–54, December 2006.

[5] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *ICML*, 2004.

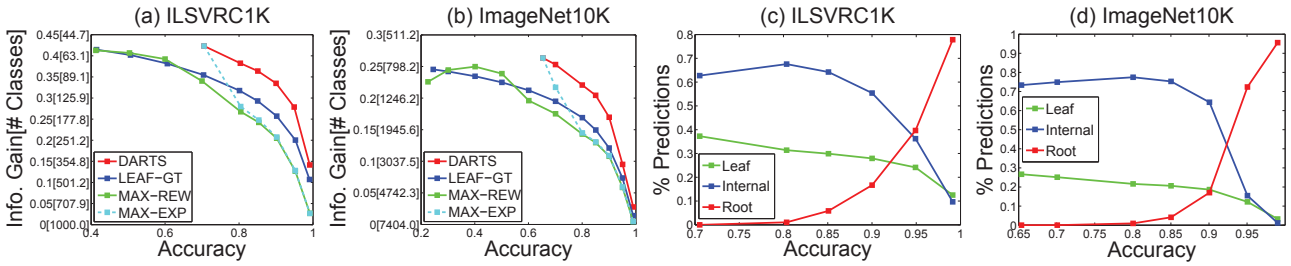

Figure 6. Large scale results: reward-vs-accuracy curves and distributions of predictions. Here the "internal nodes" exclude the root node.

.

Flat	bobsled	pheasant	mortar	canoe	loggerhead	cannon	Bouvier des Flandres	grapefruit
Ours	vehicle	animal	edible fruit	watercraft	animal	animal	living thing	citrus fruit

Figure 7. Predictions of "unusual" images by the flat classifier versus ours with a .7 accuracy guarantee, both trained on ILSVRC1K.

[6] J. J. del Coz and A. Bahamonde. Learning nondeterministic classifiers. *JMLR*, 10:2273–2293, December 2009.

[7] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10, 000 image categories tell us? In *ECCV*, 2010.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] J. Deng, S. Satheesh, A. C. Berg, and L. Fei-Fei. Fast and balanced: Efficient label tree learning for large scale object recognition. In *NIPS*, 2011.

[10] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *JMLR*, 11:1605–1641, 2010.

[11] H. Everett. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11(3):399–417, 1963.

[12] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, 2010.

[13] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.

[14] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[15] E. Grall-Maes and P. Beauseroy. Optimal decision rule with class-selective rejection and performance constraints. *PAMI*, 31(11):2073 –2082, 2009.

[16] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. *CVPR*, 2008.

[17] T. Ha. The optimum class-selective rejection rule. *PAMI*, 19(6):608 –615, Jun 1997.

[18] S. J. Hwang, K. Grauman, and F. Sha. Learning a tree of metrics with disjoint visual features. In *NIPS*, 2011.

[19] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *CVPR*, 2011.

[20] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007.

[21] M. Marszalek and C. Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, 2008.

[22] H. Masnadi-shirazi and N. Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive svms. In *ICML*, 2010.

[23] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[24] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[25] J. Platt. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 2000.

[26] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, 2011.

[27] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30(11):1958–1970, November 2008.

[28] `http://www.image-net.org/challenges/LSVRC/2010/`.

[29] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. Zhang. Content-based hierarchical classification of vacation images. In *ICMCS, Vol. 1*, pages 518–523, 1999.

[30] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[31] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[32] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *JMLR*, 11:111–130, March 2010.

[33] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, 2002.

[34] B. Zhao, L. Fei-Fei, and E. Xing. Large-scale category structure aware image categorization. In *NIPS*, 2011.

[35] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007.

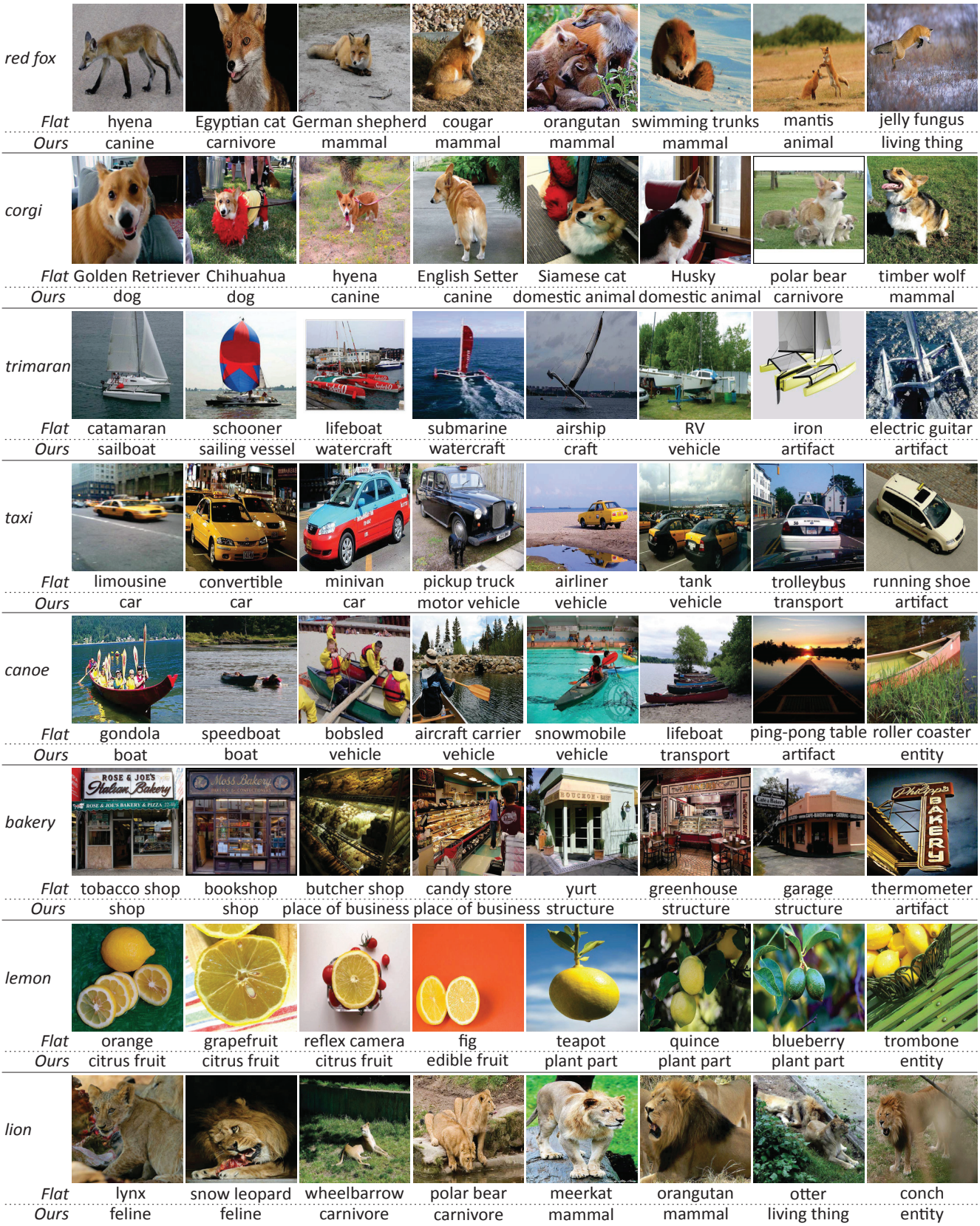

Figure 9. "Hard" test images in ILSVRC1K and the predictions made by a flat classifier and our classifier with a .8 accuracy guarantee. The flat classifier makes mistakes whereas ours stays accurate by "hedging its bets".