Level Selector Network for Optimizing Accuracy-Specificity Trade-offs

Ahsan Iqbal Juergen Gall
University of Bonn
{iqbalm, gall}@iai.uni-bonn.de

Abstract

With the increase in visual categories that become more and more fine-granular, maintaining high accuracy is a challenge. As the visual world can be organized in a semantic hierarchy, which is usually in form of a directed acyclic graph of many levels of abstraction, a classifier should be able to select an appropriate level trading off specificity for accuracy in case of uncertainty. In this work, we study the problem of finding accuracy vs. specificity trade-offs. To this end, we propose a Level Selector network, which selects the class granularity for the class prediction for an image or video, and a self-supervision based training strategy to train the Level Selector network. We show as part of the empirical evaluation, that our approach achieves superior results compared to the current state of the art on large-scale image and video datasets.

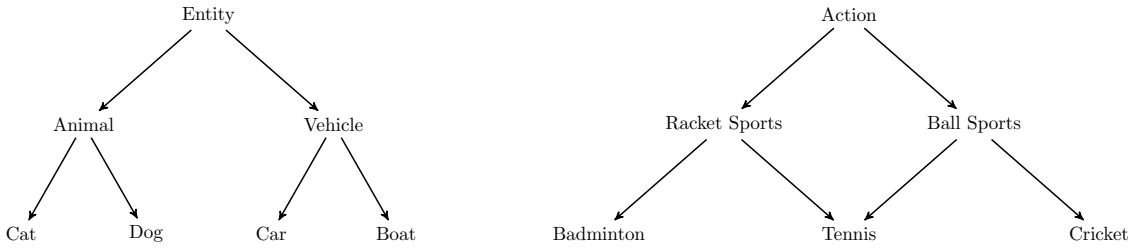
1. Introduction

Classification is a fundamental task in computer vision or machine learning and it is a building block of many other computer vision tasks such as object detection, scene parsing, or action localization. Over the years, the datasets for image or action classification not only increased in the number of samples but also in the number of classes [7, 15]. With an increasing number of target classes, however, also the granularity of the classes becomes finer and maintaining a high classification accuracy becomes a challenge. One way to address this issue is to exploit a semantic hierarchy that organizes the visual world as shown in Figure 1. For example, a Siberian Husky is also a Spitz, a dog, an animal and an entity. While it is correct to label an image of a Siberian Husky as Siberian Husky, Spitz, dog, animal or entity, the specificity of the labels varies. Although the label with highest specificity is preferred when the prediction is correct, it might be preferable to predict a label with less specificity in case of uncertainty. For instance, correctly classifying an image as dog is often better than classifying the image wrongly as Alaskan Malamute. This requires to find an optimal trade-off between accuracy and specificity. In the two extreme cases, either the label with highest specificity is always predicted with a high risk of classification errors or the entity label is predicted all the time, which is always correct but not specific.

The first approach that addresses trade-offs between accuracy and specificity has been proposed by Deng et al. [8]. The so-called Dual Accuracy Reward Trade-off Search (DARTS) is an optimization approach for finding the optimal trade-off between accuracy and specificity. It selects for a given image and the estimated classification probabilities for the classes with highest specificity, the label at a level in the hierarchy where the prediction is one hand certain and on the other hand as specific as possible. In this work, we revisit the problem but we propose a learning instead of an optimization approach to find trade-offs between accuracy and specificity. To this end, we propose a Level Selector network that selects an appropriate level of the class hierarchy given an image or video. We first use an off-the-shelf network for image classification or action recognition which is trained to predict the class probabilities for the leaf nodes in a directed acyclic graph, which represents the semantic hierarchy. The Level Selector network then takes these probabilities as input and selects the level of the hierarchy. From the selected level, the class with the highest probability is taken. In order to train the Level Selector network for an accuracy-specificity trade-off, we use a combination of a cross-entropy and ranking loss. In the experimental evaluation, we show that the proposed approach outperforms DARTS [8].

2. Related Work

Our work is related to hierarchical classification or multi-class classification [5, 10, 14, 18, 9, 4, 1, 11, 17, 12]. In [12] authors proposed a method of generating generic descriptions of videos based on semantic hierarchies of subjects, objects and verbs. In [5] the authors combine the ideas of large margin kernel methods and Bayesian analysis to do hierarchical classification. The authors of [14] proposed a method to do visual concept learning, where concepts are arranged in a hierarchy. The authors of [4] proposed a hierarchical loss to learn the paths in the class hierarchy. Accu-



(a) Class hierarchy arranged in a tree.

(b) Class hierarchy arranged in a directed acyclic graph.

Figure 1: Examples of class hierarchies. (a) Subset of classes from ILSVRC65 arranged in a tree. (b) Subset of classes from Kinetics arranged in a directed acyclic graph.

racy vs. efficiency in multi-class classification is addressed in [1, 9, 11]. They formulated hierarchical classification as a multi-class classification problem. These methods, however, do not address the issue of automatically selecting an appropriate level in the class hierarchy given an image or video. The approaches [6, 16] deal with the problem when the training samples are not always annotated at the finest granularity as it is required for testing. Instead, they investigate how training samples that are annotated at a coarser level can be used for training as well.

The method proposed in Deng et al. [8] is closely related to our approach, as it addresses the same problem of finding trade-offs between accuracy and specificity. They proposed a Dual Accuracy Reward Trade-off Search (DARTS) algorithm. The DARTS expects the probability distribution at the leaf level, an accuracy guarantee and the taxonomy as an input, and automatically selects an appropriate level in the taxonomy for a given input by optimizing the accuracy vs. specificity trade-off. In contrast to DART, we do not use an optimization approach to find the trade-offs, but we propose a network that selects the level in the hierarchy for a given image or video.

3. Accuracy vs. Specificity

We propose an approach that classifies a given video or image x, but the specificity of the predicted label varies depending on the uncertainty. We assume that a class taxonomy is given in form of a tree or directed acyclic graph as in Figure 1. The goal is then to predict the label as specific as possible while being constrained on a correct classification.

The trade-off can be formulated as in [8]. We assume that the class hierarchy is arranged in a directed acyclic graph $G = (V, E)$, with a unique root node, as in Figure 1. Each node $v \in V$ represents a semantic class and a directed edge $(P, C) \in E$ represents a parent-child relationship between parent P and child C. The leaf nodes $Y \subset V$ are mutually exclusive. Given such a hierarchy, it is then correct to label the input either by its ground truth leaf node or any of the ancestors of the ground truth leaf node. Hence, the accuracy of such a classifier for N images or videos can be defined as

$$\phi(f) = \frac{1}{N} \sum_{i=1}^{N} [f(x_i) \in \pi(y_i)] \tag{1}$$

where $[x]$ is an indicator function, which is 1 if x is true and 0 otherwise. For each image x_i, $f(x_i) \in V$ denotes the predicted node by the classifier, $y_i \in Y$ is the ground truth class at the leaf level and $\pi(y_i)$ is the set of correct labels for input x_i, i.e. the ground truth leaf node and its ancestors in the class hierarchy.

As the goal of a classifier is to achieve high accuracy, always predicting the root node will result in 100% accuracy, which results in an uninformative solution. Hence, it is preferred for a classifier to be as specific as possible. This preference can be encoded by the information gain. The information gain is defined as the expected reduction in the entropy from the prior distribution. Hence, the information gain for predicting a node v is

$$IG(f) = \log(|Y|) - \frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{y \in Y} [f(x_i) \in \pi(y)] \right) \tag{2}$$

where Y represents the set of leaf nodes in the class hierarchy. The information gain will be zero if $f(x_i)$ is always the root node since the root node is an ancestor of all leaf nodes, and it will be maximized if always leaf nodes are predicted. In the latter case, $IG(f) = \log(|Y|)$. We therefore normalise the information gain

$$IGN(f) = \frac{IG(f)}{\log(|Y|)}. \tag{3}$$

such that $IGN(f) \in [0, 1]$.

Since the optimal trade-off between accuracy and specificity depends on the application, we will measure the performance by learning the network, which will be described next, for different trade-offs. As in [8], we plot the normalised information gain (3) with respect to the accuracy (1).

4. Level Selector Network

In order to obtain various accuracy and specificity trade-offs, we propose a Level Selector network which selects an appropriate level in the semantic hierarchy for a given image or video. In a directed acyclic graph, the level of a node is defined by the minimum number of nodes that need to be passed to reach the root node, i.e. the level is zero for the root node itself, one for the children of the root node, and so on.

The Level Selector network expects a probability distribution for each level of the class hierarchy as input and outputs the probability of selecting each level. In order to train the Level Selector network, a flat classifier trained for all leaf level classes is converted into a hierarchical one. The leaf node probabilities are obtained by the flat classifier, and the probability for each internal node v can be computed by summing up the probabilities of its children by

$$P_v = \sum_{y \in Y} [v \in \pi(y)] P_y \tag{4}$$

where Y is the set of all leaf nodes. Once we have probabilities at each node in the class hierarchy, we train the Level Selector network to select a level in the class hierarchy given an input x.

The Level selector network is trained using self-supervision to maximize the information gain objective. The Level Selector takes the probability distribution at each level of the class hierarchy as input. It has H neurons in its output layer, where H is the height of the class hierarchy or the number of levels in the class hierarchy. The output of the ith neuron will be the probability of selecting the ith level in the class hierarchy for a given input. Figure 2 shows an example of a level selector network for a class hierarchy of 2 levels and a root node.

4.1. Training Strategy

The Level Selector network is trained using self-supervision. Given a training image or video x_i, we first determine the target level h_i that should be selected by the Level Selector network. To this end, we first use an off-the-shelf classifier to obtain the class probabilities for the classes at the leaf nodes. The probabilities for all nodes V are then computed as in (4). As described in Algorithm 1, we start at the leaves, i.e. $h = H$, and move to a lower level until the prediction is correct. A prediction at a level h is given by $v = argmax_{v \in V_h} P_v$, where V_h denotes all nodes at the level h, and the prediction is correct if $v \in \pi(y_i)$, where y_i is the ground-truth label for the training sample x_i at the leaf level. Note that there can be more than one correct label at a particular level h in a directed acyclic graph.

Algorithm 1 Target Calculator

```
Given x_i compute P_v using (4)
h ← H
h_i ← 0
for h ≥ 0 do
    v = argmax_{v∈V_h} P_v
    if v ∈ π(y_i) then
        h_i ← h
        break
    h ← h − 1
return h_i
```

We train the Level Selector network using the cross entropy loss

$$CE(x_i, y_i) = -\log\left(f(x_i)_{h_i}\right) \tag{5}$$

where $f(x_i)_{h_i}$ denotes the predicted probability of the network for the target level h_i. In order to encode the trade-off between accuracy and information gain, we add a ranking loss [2]:

$$RL(x_i) = -\sum_{h=1}^{H} \lambda \log\left(r_h(x_i)\right) + (1 - \lambda) \log\left(1 - r_h(x_i)\right) \tag{6}$$

where $r_h(x_i)$ is defined by

$$r_h(x_i) = \frac{1}{1 + \exp\left(f(x_i)_{h-1} - f(x_i)_h\right)}. \tag{7}$$

If $r_h(x_i)$ is larger than 0.5, it means that $f(x_i)_h$ is larger than $f(x_i)_{h-1}$ and therefore layer h is preferred to layer $h - 1$. If $r_h(x_i)$ is smaller than 0.5, it is the other way around.

By combining the two loss functions, we obtain

$$L(x_i, y_i) = CE(x_i, y_i) + RL(x_i), \tag{8}$$

where the parameter λ (6) steers the trade-off between accuracy and specificity. If $\lambda = 1$, the specificity is preferred to accuracy during training since we minimise $-r_h(x_i)$ for all h. If λ is decreased the trade-off shifts towards the accuracy.

5. Experiments

For empirical evaluation, we use three large scale datasets namely ILSVRC65, ILSVRC1K and Kinetics [15].

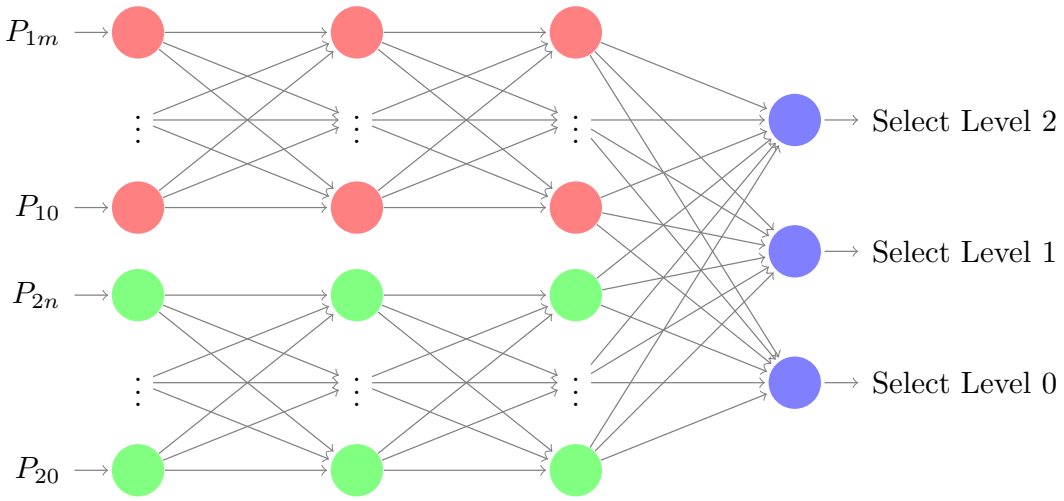


Figure 2: Level Selector Neural Network for a class hierarchy of height 3, $\{P_{2i}\}_{i=0}^{n}$, $\{P_{1i}\}_{i=0}^{m}$ are probability distributions at level 2 and level 1, obtained from an off-the-shelf classifier trained for the classes at the leaf nodes.

Dataset	Leaf Nodes	Internal Nodes	Height
ILSVRC65	57	8	4
ILSVRC1K	1000	860	19
Kinetics	400	40	3

Table 1: Dataset statistics

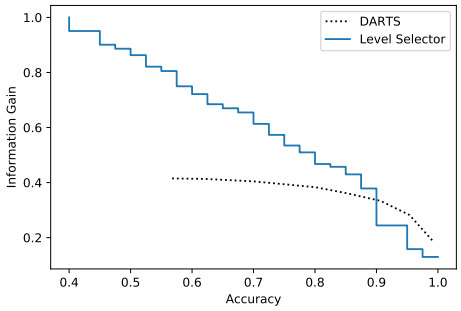
While ILSVRC65 and ILSVRC1K are subsets of ImageNet [7], which is a large scale image classification dataset where the classes are arranged by the WorldNet hierarchy, Kinetics is a large scale action recognition dataset, where the classes are arranged by a directed acyclic graph as illustrated in Figure 1. The videos in Kinetics are 10 seconds long and cut from YouTube videos. Table 1 list the class statistics of the datasets.

ILSVRC65:

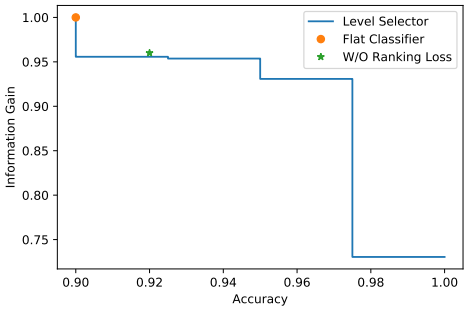
We first report the results on ILSVRC65 since the dataset has been used by Deng et al. [8] for evaluation, which is the only approach that addressed the problem of finding trade-offs between accuracy and specificity so far. For a fair comparison, we use the class probabilities for each node in the semantic hierarchy, which are provided by Deng et al. [8]. Figure 3 (a) compares the proposed Level Selector with DARTS [8]. As discussed in Section 3, the plot shows the curves for different accuracy-specificity trade-offs where specificity is measured by the normalised information gain. It can be observed that the proposed approach achieves better trade-offs. The proposed approach also achieves the full range for information gain, while DARTS gets stuck at a normalised information gain of about 0.4.

Since the classifier that has been used in [8] to obtain the class probabilities is based on SVMs, we also report the results of a more recent approach, a convolutional neural network is used for predicting the class probabilities. To this end, we use a ResNet152 [13] as off-the-shelf image classifier. The plot in Figure 3 (b) shows that ResNet improves the accuracy-specificity trade-offs. While in Figure 3 (a) the maximum information gain is achieved at an accuracy of around 0.4, the maximum information gain is already achieved at an accuracy of around 0.9 for ResNet.
We also include the results for using only a flat classifier without level selection and the Level Selector network without the ranking loss. Note that both approaches do not offer the possibility to adjust the trade-off between accuracy and specificity.

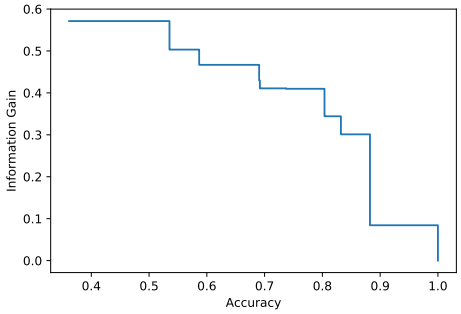
Figure 4 shows for a few examples the predictions of the proposed approach and DART. For the comparison, we used a trade-off where both approaches achieve nearly the same accuracy. The qualitative results show that the proposed method tends to be more specific. As it is an issue with most large scale datasets, ImageNet [7] contains also several annotations errors. In particular, species are not always correctly annotated. For instance, the second bird in the third row in Figure 4 is wrongly annotated as black grouse. While DARTS classifies the bird wrongly as prairie chicken, the proposed predicts correctly that it is an animal, but labeling it as bird would be more specific.



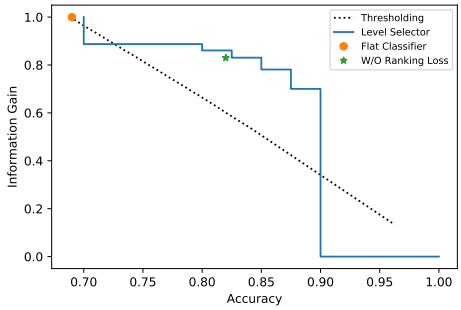
(a) Comparison of the proposed approach with DARTS [8] on ILSVRC65 using the class probabilities from [8].



(b) Accuracy vs. information gain using ResNet as off-the-shelf classifier on ILSVRC65.



(c) Accuracy vs. information gain using ResNet as off-the-shelf classifier on ILSVRC1K



(d) Accuracy vs. information gain using I3D as off-the-shelf classifier on Kinetics

Figure 3: Accuracy vs. information gain trade-offs

ILSVRC1K:

We also evaluate the approach on the larger subset ILSVRC1K. In contrast to ILSVRC65, the semantic hierarchy is not represented by a tree but by a directed acyclic graph. Hence, a node can have more than one parent. As off-the-shelf classifier, we use ResNet152 as before. The accuracy-specificity plot is shown in Figure 3 (c). In contrast to ILSVRC65, the proposed network does not achieve the full information gain, but it gets stuck slightly below 0.6. This can be explained by the large number of levels in the semantic hierarchy as reported in Table 1 and the noisy annotations at the leaf nodes.

Kinetics:

We finally evaluate the approach on Kinetics, which is a dataset for action recognition. As off-the-shelf classifier, we use the I3D network [3], which takes the RGB frames and optical flow as input. Since the semantic hierarchy of Kinetics has only three layers, we can compare the approach with a baseline that uses a threshold to steer the accuracy-specificity trade-off. Given the class probabilities for the leaf nodes, we take the class with highest probability if it is above a threshold. If it is below the threshold, we take the class with highest probability at the second level. If this probability is also below the threshold, we select the root. Figure 3 (d) shows that the proposed approach provides better trade-offs than the thresholding baseline. In particular for an accuracy between 0.7 and 0.9, the information gain is very high, which shows the high specificity for this accuracy range.

6. Conclusion

We revisited the problem of finding accuracy-specificity trade-offs for image classification or action recognition. Instead of an optimization based approach, we proposed a network that learns to select the right granularity based on the class probabilities of an off-the-shelf classifier. We showed that the network learns various trade-offs ranging from 100% accuracy but low specificity to high specificity but moderate accuracy and that it achieves better trade-offs than previous work.

Acknowledgement The work has been funded by

GT: Race Car
DARTS: Car
LS: Race Car

GT: Race Car
DARTS: Race Car
LS: Race Car

GT: Wagon
DARTS: Car
LS: Wagon

GT: Convertible
DARTS: Car
LS: Convertible

GT: Fireboat
DARTS: Boat
LS: Fireboat

GT: Canoe
DARTS: Boat
LS: Vehicle

GT: Fireboat
DARTS: Fireboat
LS: Fireboat

GT: Speedboat
DARTS: Boat
LS: Speedboat

GT: Partridge
DARTS: Partridge
LS: Partridge

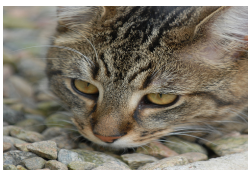
GT: Black Grouse
DARTS: Prairie Chicken
LS: Animal

GT: Pheasant
DARTS: Bird
LS: Pheasant

GT: Quail
DARTS: Quail
LS: Quail

GT: Egyptian Cat
DARTS: House Cat
LS: Egyptian Cat

GT: Tiger Cat
DARTS: Tiger Cat
LS: Tiger Cat



GT: Tiger Cat
DARTS: House Cat
LS: Tiger Cat

GT: Siamese
DARTS: House Cat
LS: Siamese

GT: Australian terrier
DARTS: Dog
LS: Dog

GT: Australian terrier
DARTS: House Cat
LS: Animal

GT: St Bernard
DARTS: Dog
LS: St Bernard

GT: German shepherd
DARTS: German shepherd
LS: German shepherd

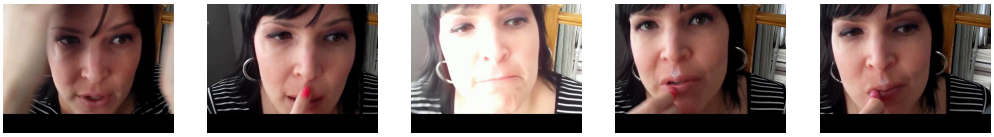
Figure 4: Qualitative results on ILSVRC65. For nearly the same recognition accuracy, the proposed Level Selector is more specific than DARTS.

GT Fine: playing cricket
GT Coarse: ball sports, racket + bat sports
Predicted Fine: playing cricket
Predicted Coarse: ball sports
Predicted Level: Fine

GT Fine: abseiling
GT Coarse: heights
Predicted Fine: abseiling
Predicted Coarse: heights
Predicted Level: Fine

GT Fine: baking cookies
GT Coarse: cooking
Predicted Fine: barbequing
Predicted Coarse: cooking
Predicted Level: Coarse

GT Fine: barbequing
GT Coarse: cooking
Predicted Fine: making a sandwich
Predicted Coarse: cooking
Predicted Level: Coarse



GT Fine: applying cream
GT Coarse: makeup
Predicted Fine: brushing teeth
Predicted Coarse: head + mouth
Predicted Level: some action (uncertain)

Figure 5: Qualitative results on Kinetics.

the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/4-1 (FOR 2535 Anticipating Human Behavior) and the ERC Starting Grant ARCA (677650).

References

[1] Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *Advances in Neural Information Processing Systems*, 2010.

[2] Chris Burges, Tal Shaked, Erin Renshaw, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Int. Conf. on Machine Learning*, 2005.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[4] Nicolò Cesa-bianchi, Claudio Gentile, Andrea Tironi, and Luca Zaniboni. Incremental algorithms for hierarchical classification. In *Advances in Neural Information Processing Systems*, 2005.

[5] Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Int. Conf. on Machine Learning*, 2004.

[6] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conf. on Computer Vision*, 2014.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[8] Jia Deng, Jonathan Krause, Alexander C. Berg, and Fei-Fei Li. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[9] Jia Deng, Sanjeev Satheesh, Alexander C. Berg, and Fei-Fei Li. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems*, 2011.

[10] Rob Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. Semantic label sharing for learning with many categories. In *European Conf. on Computer Vision*, 2010.

[11] Tianshi Gao and Daphne Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *Int. Conf. on Computer Vision*, 2011.

[12] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Int. Conf. on Computer Vision*, 2013.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[14] Yangqing Jia, Joshua Abbott, Joseph Austerweil, Thomas Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems*, 2013.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[16] Marko Ristin, Juergen Gall, Matthieu Guillaumin, and Luc Van Gool. From categories to subcategories: Large-scale image classification with partial class label refinement. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[17] Bin Zhao, Fei-Fei Li, and Eric P. Xing. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems*, 2011.

[18] Alon Zweig and Daphna Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *Int. Conf. on Computer Vision*, 2007.