ONE SIZE DOESN'T FIT ALL: ADAPTIVE LABEL SMOOTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper concerns the use of objectness measures to improve the calibration performance of Convolutional Neural Networks (CNNs). Objectness is a measure of likelihood of an object from *any* class being present in a given image. CNNs have proven to be very good classifiers and generally localize objects well; however, the loss functions typically used to train classification CNNs do not penalize inability to localize an object, nor do they take into account an object's relative size in the given image. During training on ImageNet-1K almost all approaches use random crops on the images and this transformation sometimes provides the CNN with background only samples with the correct label,this causes the classifiers to overfit to context. Context dependence is harmful for safety-critical applications. We present a novel approach to object localization that combines the ideas of objectness and label smoothing during training. Unlike previous methods, we compute a smoothing factor that is *adaptive* based on relative object size within an image. This causes our approach to produce confidences that are grounded in the size of the object being classified instead of relying on context to make the correct predictions. We present extensive results using ImageNet to demonstrate that CNNs trained using adaptive label smoothing are much less likely to be overconfident in their predictions, as compared to CNNs trained using hard targets. We also show qualitative results using class activation maps to illustrate the improvements.

1 INTRODUCTION

Convolutional neural networks (CNNs) have been used for addressing many computer vision problems for over 2 decades (LeCun, 1998). In particular, CNNs have shown promising results on object detection and localization tasks since 2013 (Krizhevsky et al., 2012; Russakovsky et al., 2015). Unfortunately, modern CNNs are overconfident in their predictions (Lakshminarayanan et al., 2017; Hein et al., 2019) and they suffer from reliability issues due to miscalibration (Guo et al., 2017a). Problems related to overconfidence, generalization and reliability represent a severe limitation of current CNNs for real-world applications. It is essential for these problems to be addressed, particularly for use in safety-critical applications as identified by (Amodei et al., 2016).

Regularization approaches are used for improving generalization and minimizing overfitting characteristics of CNNs. Recently, (Szegedy et al., 2016) employed label smoothing, providing soft labels that are a weighted average of the hard targets and uniform distribution over classes during training to improve learning speed and generalization performance. In the case of classification CNNs, ground-truth labels are typically provided as a one-hot representation of class probabilities. These labels consist of $0s$ and $1s$, with a single 1 indicating the pertinent class in a given label vector. Label smoothing provides labels which are a weighted average of hard labels with a uniform distribution over all classes using a fixed smoothing factor (Szegedy et al., 2016; Müller et al., 2019). Label smoothing minimizes the gap between the logits of the classes and shows improvement in learning speed and generalization; in contrast, hard targets tend to increase the values of the logits and produce overconfident predictions (Szegedy et al., 2016; Müller et al., 2019). Label smoothing and the traditional hard labels force CNNs to produce high confidence predictions even when pertinent objects are absent during training. To obtain more reliable confidence measures, we use objectness measure to derive a smoothing factor for every sample undergoing a unique scale and crop transformation in an adaptive manner. Safely deploying deep learning based models has also become a more immediate challenge (Amodei et al., 2016), as a community, apart from obtaining high accuracies, we also need

to provide reliable uncertainty measures of CNNs. By having reliable confidence measures for CNNs, we can improve precision by not acting with certainty when uncertain predictions are produced, as in the case of safety-critical systems.

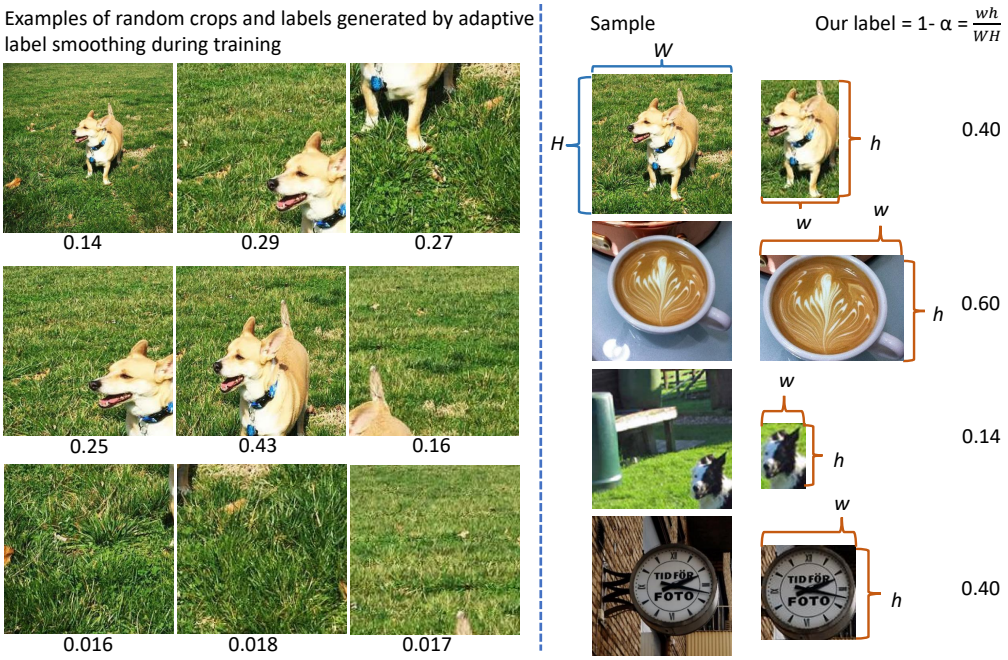


Figure 1: Random crops of images are often used when training classification CNNs to help mitigate size, position and scale bias (as shown in the left half of the figure along with the objectness values listed below them). Unfortunately, some of these crops miss the object as the process does not use any object location information. Traditional hard label and smooth label approaches do not account for the proportion of the object being classified and use a fixed label of '1' or '0.9' in the case of label smoothing. Our approach (right half) smooths the hard labels by accounting for the objectness measure to compute an *adaptive* smoothing factor. The objectness is computed using bounding box information as shown above. Our approach helps generate accurate labels during training and penalizes low-entropy (high-confidence) predictions for context-only images (the main object is completely or mostly absent).

Object detection (Girshick et al., 2018) is a well-studied problem and most approaches need bounding box information during training. Recently, (Dvornik et al., 2018) proposed using novel synthetic images to improve the object detection performance by augmenting training data using object location information. However, classification CNNs have not exploited bounding box information to regularize CNNs on large datasets to our knowledge. Objectness was first introduced by (Alexe et al., 2012), and the role of objectness has been studied extensively since then. Quantifying the likelihood an image window contains an object belonging to any class makes the measure class agnostic. Object detectors specialize in a few classes, but objectness is class agnostic by definition. We limit the definition of objectness to the 1000 ImageNet-1K classes. In other words, any object other than the objects belonging to the 1000 classes will have a score of '0'.

When training a classifier, cross entropy loss is employed but it does not penalize incorrect spatial attention, making CNNs often overfit to context or texture rather than the pertinent object (Geirhos et al., 2019), as shown in the left half of figure 1. The bottom row displays samples with negligible amounts of 'Dog' pixels and traditional methods would label them as 'Dog', causing CNNs to output incorrect predictions with a high confidence when presented with images of backgrounds or just context. Adaptive label smoothing (our approach) involves using bounding box information to smooth the hard labels of a classifier, as displayed to the right in figure 1. Our approach adapts label smoothing by deriving the smoothing factor using the objectness measure. When compared to approaches based on hard labels, sample mixing and label smoothing, our approach improves object detection and calibration performance.

Traditional approaches (Yun et al., 2019; Takahashi et al., 2018; Krizhevsky et al., 2012; Russakovsky et al., 2015) use random resize and random crop augmentation, and sometimes lose the pertinent object in the training sample, allowing the classifier to make the correct predictions by overfitting to the context surrounding the pictures. We believe that our approach addresses significant problems that are associated with current training techniques. In particular, random cropping of images is a common augmentation technique during training of ResNet, but occasionally the crop misses the object entirely. In such a case, the equivalent of a one-hot label is typically provided, with the result that the system is steered toward increased dependence on background (context) portions of the image. We argue that one-hot representations are too limiting, and our adaptive approach to label smoothing makes it possible for the classifier to avoid overconfidence in many cases. Specifically, our approach accomplishes the following: 1) Our labels not only indicate the presence of an object but also tell the classifier the gross proportion of the object in a given image. This implicit regularization guides the classifiers to avoid producing high confidence values when the object pixels are lower in proportion. On the other hand, if a random crop contains mostly object pixels, then the classifier will be encouraged to produce higher-confidence predictions. 2) A classifier trained with hard labels will tend to generate predictions (correct and wrong) with high confidence values even when images containing only background or images from a different distribution (no pertinent objects) are presented. As classifiers often produce overconfident predictions, it poses a problem for safety critical applications. With our approach, the system is trained to produce lower confidence predictions with out-of-distribution samples or background-only images are presented. Low confidence predictions from our approach are meaningful as they can easily be thresholded for better noise rejection. High confidence approaches are hard to threshold as most predictions have high confidence even when they are wrong. 3) Traditional classifiers "cheat" by relying heavily on context. Although context helps increase computed accuracy for a given dataset, such reliance is not viable for real-world applications. During training, we assume that every class is equiprobable when only background or context is provided as input.

Our approach uses labels that are more accurate than other baselines when random cropping and scaling of images are applied during training. To our knowledge, almost all classifiers trained on ImageNet use random crop and scaling based augmentation to regularize. The random crop transformation allows the CNNs to cheat by relying on context rather than the pertinent object, our approach uses bounding box labels to produce labels in an adaptive way during training. To quantify context dependence, we used bounding box annotations on the 50K validation images, removed all objects and replaced the pixels with the mean image pixel values. Hard label approach had an accuracy of 6.3% with an average confidence of 0.29, label smoothing predicted with an accuracy of 6.1% and an average confidence of 0.2, CutMix had an accuracy of 9.2% with an average confidence of 0.2. These baseline methods produced high confidence predictions on images with no objects present using just context information. Our approach had an accuracy of 4.7% and an average confidence of 0.02. We have an order of magnitude improvement in performance over recent baselines as our approach helps CNNs produce confidence based on the relative size of the pertinent object.

The main contributions of this work are as follows. 1) We introduce the concept of *adaptive label smoothing*, which is an alternative to the use of traditional hard labels during CNN training. This technique automatically adjusts labels during training based on relative size of an object in the image. This adjustment is important because images with no pertinent objects should not produce high confidences. 2) We demonstrate that objectness measures can be incorporated into adaptive label smoothing as an effective method of reducing overconfidence. Taken together, these contributions result in improved classification performance by CNNs with less likelihood of overconfidence. We have trained 13 classifiers and evaluated them on three popular datasets, and results show that for context-only images our approach produces an average confidence that is an order of magnitude lower compared to baselines. Confidence values generated by CNNs help us understand the output predictions, but unreliable confidence measures hurt the applicability of CNNs for safety-critical applications.

2 Related Work

Bias exhibited by machine learning models can be attributed to many underlying statistics present in datasets and model architectures (Battaglia et al., 2018; Zhang et al., 2017b) including context,

object texture (Geirhos et al., 2019), size, shape and color in the case of images. Various approaches to mitigate bias have been proposed (Anne Hendricks et al., 2018; Choi et al., 2019; Geirhos et al., 2019) in recent years. Our approach produces high entropy predictictions when context-only images are provided as input during inference, as we aim to learn the size of the relevant object within the image and classify it, instead of relying on context to produce a prediction.

Traditionally, any label preserving transformation on an input image is considered to help regularize a CNN. The authors of AlexNet (Krizhevsky et al., 2012) employed random cropping and horizontal flipping methods when they surpassed the performance of conventional machine learning approaches in 2012. Randomly cropping a given image during training prevents overfitting the scale or location of the object; flipping an image improves the generalization to view points. The random noise class of data augmentation methods (DeVries & Taylor, 2017; Zhong et al., 2017) mask random regions of an input image with zeros. Random noise based methods may accidentally erase the pertinent object in a given image and force the CNN to rely purely on context to make a prediction, contributing to label noise. The authors of DropBlock (Ghiasi et al., 2018) have used dropout in the feature space to obtain better generalization. Authors of AutoAugment (Cubuk et al., 2019) used reinforcement learning dynamically during training to learn the best combination of existing data augmentation methods.

In contrast to augmentation based approaches that manipulate the input but not the corresponding label, our approach regularizes classification CNNs by computing a label based on the proportion of the object being classified in a given random crop of the training sample. The latest work in the area of data augmentation uses samples from different classes and changes expected outputs to predict a probability distribution based on the number and intensity of pixels represented by each class. The authors of Mixup (Zhang et al., 2017a; Tokozume et al., 2018) use alpha blending (weighted sum of pixels from two different classes), and apply blending weights to corresponding labels.The authors of CutMix and RICAP (Yun et al., 2019; Takahashi et al., 2018) also use soft labels by cropping different regions and classes of images and 'mixing' the labels in matching proportions to corresponding regions in the final augmented sample. The sample mixing based approaches above do not rely on object size when 'mixing' regions in images and computing the label. Conversely, our approach uses bounding box information to apply a smoothing factor based on the object's size relative to the image size to produce a soft label without mixing the samples.

Calibration of CNNs is important as predictions need to be equally accurate and confident. Calibration and uncertainty estimation of predictors has been an ongoing interest to the machine learning community Murphy (1973); DeGroot & Fienberg (1983); Platt (1999); Lin et al. (2007); Zadrozny & Elkan (2002). Bayesian binning into quantiles (BBQ) Naeini et al. (2015) was proposed for binary classification and beta calibration, Kull et al. (2017) employed logistic calibration for binary classifiers. In the context of CNNs, Guo et al. (2017b) proposed a temperature scaling approach to improve calibration performance of pre-trained models. Calibration has been explored in multiple directions; popular approaches to calibrate CNNs are to transform outputs of pre-trained models using approximate bayesian inference Maddox et al. (2019), or to use a special loss function to help regularize the model (Pereyra et al., 2017; Kumar et al., 2018) during training. Our approach is loosely related to the latter class of methods. Our work also relates to label smoothing proposed first by (Szegedy et al., 2016), with its applicability for many tasks explored by (Pereyra et al., 2017); (Xie et al., 2016) applies dropout-like noise to the labels. Recently, (Müller et al., 2019) explored the benefits of label smoothing; apart from having a regularizing effect, label smoothing helps reduce intra-class distance between samples (Müller et al., 2019). Another approach to calibrate CNNs was proposed by (Mukhoti et al., 2020), using a special loss function and temperature scaling, the authors were able to obtain state-of-the-art calibration performance. Label smoothing also improves calibration performance of CNNs (Mukhoti et al., 2020).

Contrasting previously discussed methods, our approach involves using hard labels multiplied by the objectness measure and obtaining a uniform distribution over all other classes when input images are devoid of pertinent objects. We do not change our loss function as opposed to (Mukhoti et al., 2020). Our approach can be described as a variant of label smoothing, employing an *adaptive* label smoothing approach that is unique to every training sample as it accounts for object size. To our knowledge, we are the first to apply *adaptive* label smoothing to train image classification CNNs.

The objectness is computed using bounding box information during training. CNNs trained using hard labels produce 'peaky' probability distributions without considering the spatial size of the

pertaining object. Our approach produces outputs that are softer and the peaks correspond to the spatial footprint of the object being classified as shown in figure 3.

3 METHOD

Consider $D = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ to be a dataset consisting of N independent and identically distributed real-world images belonging to K different classes. Let $\mathcal{X}$ represent the set of images, and let $\mathcal{Y}$ denote the set of ground-truth class labels. Sample i consists of the image $\mathbf{x}_i \in \mathcal{X}$ along with its corresponding label $y_i \in \mathcal{Y} = \{1, 2, ..., K\}$. Let f_θ represent the CNN classifier with model parameters θ. The predicted class is $\hat{y}_i = \operatorname{argmax}_{y \in \mathcal{Y}} \hat{p}_{i,y}$, where $\hat{p}_{i,y} = f_\theta(y|\mathbf{x}_i)$ is the computed probability that image $\mathbf{x}_i$ belongs to the class y.

Let z_i represent the one-hot encoding of label y_i. Following (Szegedy et al., 2016), the hard label z_i can be converted to soft label $\tilde{z}_i$ using $\tilde{z}_i = z_i(1-\alpha) + (1-z_i)\alpha/(K-1)$, where $\alpha \in [0, 1]$ is a fixed hyperparameter. This is the standard procedure known as label smoothing.

A novelty of our approach is to make α *adaptive*, calculating the value based on the relative size of an object within a given training image. Using the bounding box annotations available for the images in the dataset, we generate object masks. We apply the same augmentation transform (scale, crop) to the masks and compute the objectness score on the fly for every training image. Let the image width and height be denoted by (W, H) and the object width and height be denoted by (w, h). The ratio α is computed as $\alpha = 1 - \frac{wh}{WH}$. The soft label $\tilde{z}_i$ is computed as before. We also explore a weighted combination of *adaptive* label smoothing and hard labels. To do this, we introduce parameter $\beta \in [0, 1]$ to determine the degree of *adaptive* label smoothing being applied. The setting $\beta = 0$ corresponds to the case of classic hard labels. The soft label in this case is computed as $\tilde{z}_i = (z_i(1-\alpha) + (1-z_i)\alpha/(K-1))\beta + (1-\beta)(z_i)$.

4 EXPERIMENTS

In this section, we provide a description of the datasets used in our experiments, introduce some of the commonly used metrics for calibration of CNNs and describe our implementation details. We then discuss the merits of our approach and answer important questions related to applicability to transfer learning in an object detection setting, and we discuss the effect of using different types of labels during training in an ablative manner. We use ResNet-50 (He et al., 2016) for all our experiments.

4.1 DATASETS

As indicated in table 1, we have used different training datasets that are based on ImageNet-1K dataset (Russakovsky et al., 2015). ImageNet-1K consists of 1.2M training images and 50K validation images spanning 1K categories. As only 38% of ImageNet training images have bounding-box annotations, we distinguish these experiments from those trained on the full dataset. We use standard data-augmentation strategies for all methods and train all our models for 300 epochs starting with a learning rate of 0.1 and decayed by 0.1 at epochs 75, 150, and 225 using a batch size of 256. As shown in table 1, we have different training datasets that are based on ImageNet-1K dataset (Russakovsky et al., 2015).

The first 2 rows of the table employ the standard dataset for training. However, as our method needs object proportions, we use a subset of the standard ImageNet dataset that has bounding boxes (0.474M). These results are shown in the next 8 rows of table 1. To generate the 'mask' version, we make sure that only one object is present in a given image and 'mask' all other objects replacing them with pixel means. We use this version of the dataset derived from the 0.474M subset and identify the approach with '(mask)' next to the method in table 1. We end up with about 54K more images as some ImageNet images have multiple annotated objects. Lastly, we generate another dataset that is devoid of any object altogether. We sample about 15% of the time from this dataset during training of one of our approaches, and the label generated for these methods is a vector of uniform probability distribution across 1000 classes. The idea is that when no objects are present in a sample, a CNN should produce a high-entropy prediction.

For validation, we use the validation set of (Russakovsky et al., 2015) (V1) and the newly released ImageNetV2 set (Recht et al., 2019). Specifically, we use the more challenging 'MatchedFrequency' set of images. The different validation sets are identified in the 'Val.' column of table 1.

To measure the transfer-learning ability of the representations learned by our classifiers, we used the challenging (Lin et al., 2014) dataset to obtain the results described in-2. The dataset consists of about 230K training images and we use the 'minival' validation set of 5K images.

Table 1: Classification and calibration results with ImageNet. For a detailed explanation of the metrics please refer to 4.2.'O.conf' and 'U.conf' refer to overconfidence and underconfidence scores.

Method	Val. Set	Train N	ACC	ECE 100	ECE 15	MCE	O.conf	U.conf
Hard Label	V1	1.28M	0.769	0.062	0.045	0.284	0.582	0.100
Hard Label	V2	1.28M	0.647	0.131	0.099	0.664	0.538	0.131
Hard Label	V1	0.474M	0.669	0.104	0.093	0.347	0.558	0.123
Hard Label	V2	0.474M	0.543	0.171	0.148	0.759	0.520	0.154
CutMix	V1	0.474M	0.689	0.032	0.017	0.167	0.456	0.209
CutMix	V2	0.474M	0.577	0.100	0.050	0.517	0.421	0.248
Label Smoothing	V1	0.474M	0.691	0.055	0.051	0.354	0.401	0.248
Label Smoothing	V2	0.474M	0.558	0.102	0.047	0.512	0.368	0.283
A. L. S. (Ours)	V1	0.474M	0.655	0.191	0.186	0.461	0.255	0.401
A. L. S. (Ours)	V2	0.474M	0.532	0.185	0.158	0.661	0.217	0.441
Hard Label (mask)	V1	0.528M	0.680	0.088	0.076	0.259	0.549	0.132
Hard Label (mask)	V2	0.528M	0.559	0.155	0.127	0.686	0.507	0.163
CutMix (mask)	V1	0.528M	0.698	0.032	0.020	0.249	0.477	0.197
CutMix (mask)	V2	0.528M	0.576	0.110	0.067	0.614	0.449	0.228
Label Smoothing (mask)	V1	0.528M	0.687	0.051	0.046	0.430	0.407	0.244
Label Smoothing (mask)	V2	0.528M	0.563	0.108	0.048	0.524	0.374	0.281
A. L. S. (mask)	V1	0.528M	0.648	0.186	0.182	0.463	0.246	0.396
A. L. S. (mask)	V2	0.528M	0.528	0.185	0.160	0.687	0.209	0.441

Table 2: Fine-tuning on COCO using FRCNN for object detection. For a detailed explanation of the results please refer to 4.3. AP refers to average precision and AR refers to average recall at the specified Intersection over union (IoU) level.

Method	Pre-trained on	Pre-train N	AP (0.5:0.95)	AP (0.5)	AP (0.75)	AR (0.5:0.95)
Hard Label	ImageNet	1.28M	0.323	0.519	0.345	0.438
CutMix	ImageNet	1.28M	0.329	0.528	0.353	0.445
RICAP	ImageNet	1.28M	0.331	0.528	0.354	0.447
Hard Label	ImageNet	0.474M	0.290	0.479	0.309	0.415
A. L. S. (Ours)	ImageNet	0.474M	0.311	0.501	0.332	0.429
Hard Label (mask)	ImageNet	0.528M	0.290	0.482	0.307	0.415
CutMix (mask)	ImageNet	0.528M	0.312	0.509	0.329	0.428
Label Smoothing (mask)	ImageNet	0.528M	0.304	0.500	0.324	0.424
A. L. S. (mask)	ImageNet	0.528M	0.311	0.501	0.333	0.428
A. L. S. (mask) (Beta=0.75)	ImageNet	0.528M	0.309	0.498	0.331	0.427
A. L. S. (mask) (Beta=0.25)	ImageNet	0.528M	0.298	0.492	0.315	0.419
A. L. S. + Context (mask)	ImageNet	0.528M	0.303	0.490	0.323	0.421
A. L. S. + CutMix (mask)	ImageNet	0.528M	0.273	0.449	0.289	0.403

4.2 CLASSIFICATION AND CALIBRATION

This section identifies various calibration metrics used by the community and discusses our results obtained on the popular (Russakovsky et al., 2015; Recht et al., 2019) datasets. We use the implementation of (Wenger et al., 2020) on all of our classifiers to generate the results in table 1. To evaluate the performance of *adaptive* label smoothing we use five metrics that are very common: *accuracy*, *expected calibration error* (ECE) (Naeini et al., 2015), *maximum calibration error* (MCE) (Naeini et al., 2015), *overconfidence* (Mund et al., 2015), and *underconfidence* (Mund et al., 2015). We computed ECE using 100 bins and 15 bins. The authors of (Wenger et al., 2020; Kumar et al., 2019) discuss the advantages of using 100 bins in greater detail.

ECE is defined as the expected absolute difference between a classifier's confidence and its accuracy using a finite number of bins (Naeini et al., 2015). MCE is defined as the maximum absolute difference between a classifier's confidence and its accuracy of each bin (Naeini et al., 2015). Overconfidence is the average confidence of a classifier's false predictions and underconfidence as the average uncertainty on its correct predictions (Wenger et al., 2020; Mund et al., 2015).

The results in table 1 indicate our approaches based on *adaptive* label smoothing using the abbreviation 'A. L. S.' In general, these results have a low overconfidence score, which is highly desirable. These results demonstrate that

adaptive label smoothing based CNNs seldom produce high confidence scores when they make incorrect predictions. In fact, our models are underconfident as they pay attention to the spatial footprint of the pertinent object. It is important to note that our methods outperform all baselines for the overconfidence metric.

4.3 TRANSFER LEARNING FOR OBJECT DETECTION

We adopt the architecture of Faster RCNN (Ren et al., 2015) adapted to use the ResNet-50 backbone. Specifically, we train all of our classifiers using the implementation of *https://github.com/jwyang/faster-rcnn.pytorch*. We train all ImageNet pre-trained models with a batch size of 16 and initial learning rate of 0.01 decayed after every 4 epochs for a total of 10 epochs. We train all OpenImages pre-trained models with a batch size of 16 and initial learning rate of 0.0075 decayed after every 4 epochs for a total of 10 epochs. We employ the standard metrics for average precision (AP) and average recall (Lin et al., 2014) at different intersection over union (IoU) levels. As shown in 2, our approach outperforms hard label and label smoothing based approaches on this downstream task. The better localization performance is also shown without fine-tuning using class activation maps in figure 4. Specifically, our approach performs almost as well as CutMix (Yun et al., 2019) using AP measures.

4.4 ABLATION STUDIES

We compare our approach with standard baselines and provide results in an ablative manner to understand the benefits and limitations of applying *adaptive* label smoothing to classification and transfer learning for object detection tasks. As shown in figure 2, increasing the value of β helps reduce model overconfidence and produces predictions that are less 'peaky' compared to label smoothing and hard label settings. Another interesting trend can be observed by changing the value of the β parameter. As β decreases in value, the overconfidence rate goes up along with it as shown in table 1. Average confidence of a model describes the mean confidence of a model. As our model predictions are grounded in the spatial size of the object, our average confidence values on 'V1' and 'V2' are 0.48 and 0.39, respectively; in the case of hard labels the values are 0.77 and 0.69, respectively.

In case of transfer learning, we observe that decreasing β causes the object localization performance to drop. Using implicit object size information helps CNNs localize and detect objects for downstream tasks as well.

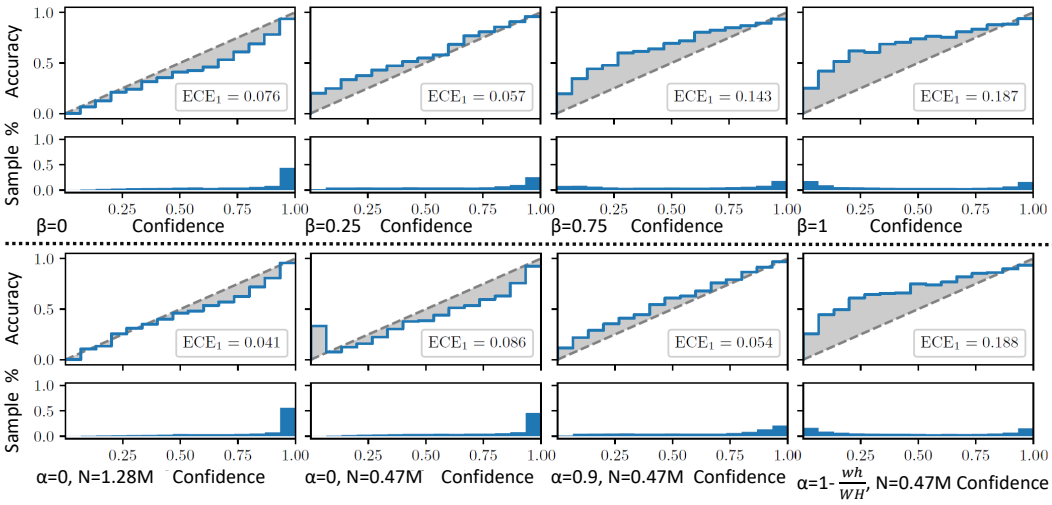


Figure 2: Reliability diagrams help understand the calibration performance (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005) of classifiers. We compute ECE_1 using the implementation of (Wenger et al., 2020) on the validation set of ImageNet. The deviation from the dashed line (shown in gray), weighted by the histogram of confidence values, is equal to Expected Calibration Error (Wenger et al., 2020). The top half of the figure shows classifiers trained using the same dataset (N=0.528M), but with different values of β. The leftmost reliability diagram is the classic hard label setting and the rightmost reliability diagram is the *adaptive* label setting. The bottom half of the figure compares classifiers trained on the complete ImageNet (leftmost) with 3 classifiers trained on the subset of ImageNet with bounding box labels using different values of the α hyperparameter.

5 Conclusion

This paper has addressed the problems of contextual bias and calibration using a novel approach called *adaptive* label smoothing. We show that bounding box information pertaining to objects can be used to compute a smoothing factor adaptively during training to improve the localization and calibration performance of CNNs. We use bounding box information for a portion of the ImageNet dataset (Russakovsky et al., 2015) to train 13 different classifiers. We show that our approach can be combined with traditional label smoothing approaches to train CNNs that are calibrated and have better localization performance on the challenging MS-COCO dataset (Lin et al., 2014) after fine-tuning, compared to approaches that use hard labels or traditional label smoothing approaches. Our labels implicitly capture the object proportion within an image during training, a significantly more challenging task than training with hard labels.

Our methods provide the lowest accuracy and an order of magnitude reduction in average confidence when presented with context only images. We are extending this work to out of distribution detection as well. With adaptive label smoothing, when no pertinent objects are present, every class is equally probable for a given image. We introduce adaptive label smoothing with the notion that safety-critical applications need CNNs that are trained not to be overconfident in their predictions. Our intention is for decision making systems (steering inputs to an autonomous vehicle for example) to not make decisions in a definite way when the models are not confident in their predictions. Our approach provides a more reliable measure of confidence compared to all baselines.

References

Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2189–2202, 2012.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 771–787, 2018.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.

Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pp. 851–863, 2019.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:12–22, 1983.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 364–380, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 10750–10760, 2018.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017a.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 623–631, 2017.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3792–3803, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2810–2819, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6402–6413, 2017.

Yann LeCun. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *arXiv preprint arXiv:1902.02476*, 2019.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4696–4705, 2019.

D. Mund, R. Triebel, and D. Cremers. Active online confidence boosting for efficient object classification. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1367–1373, 2015.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Blai Bonet and Sven Koenig (eds.), *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 2901–2907, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 625–632, 2005.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Ricap: Random image cropping and patching data augmentation for deep cnns. In *Asian Conference on Machine Learning*, pp. 786–798, 2018.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5486–5494, 2018.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Proceedings of Machine Learning Research, 2020. URL `https://github.com/JonathanWenger/pycalib`.

Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4753–4762, 2016.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6023–6032, 2019.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 694–699, 2002.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Quanshi Zhang, Wenguan Wang, and Song-Chun Zhu. Examining cnn representations with respect to dataset bias. *arXiv preprint arXiv:1710.10577*, 2017b.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

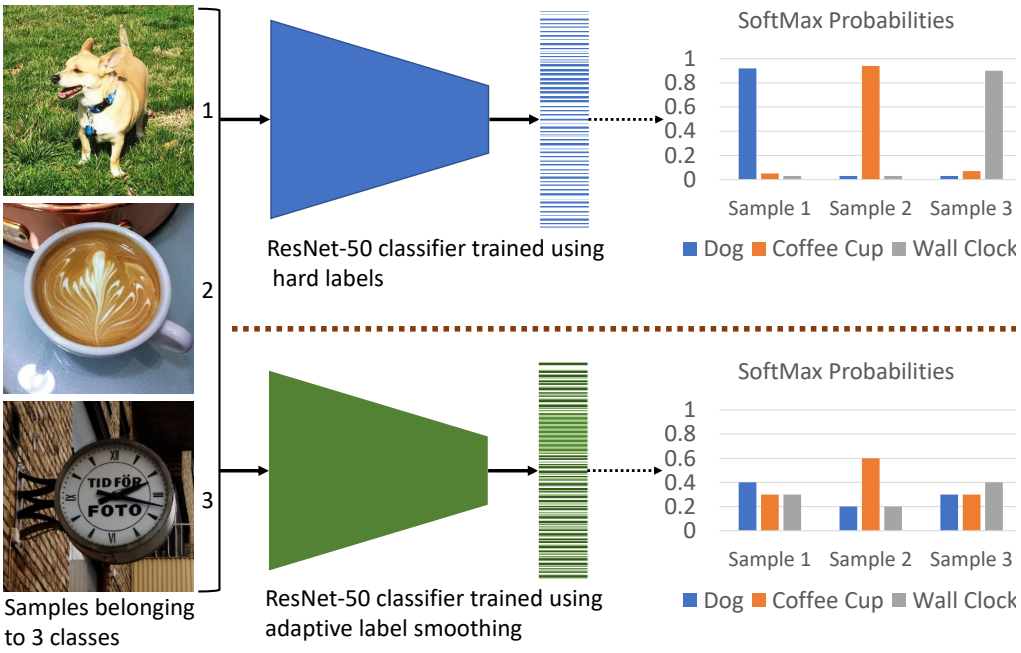


Figure 3: Hard-label and label-smoothing based approaches (top half of the figure) do not take into account the proportion of the object being classified. Our approach (bottom half) weights soft labels using the objectness measure to compute an *adaptive* smoothing factor.

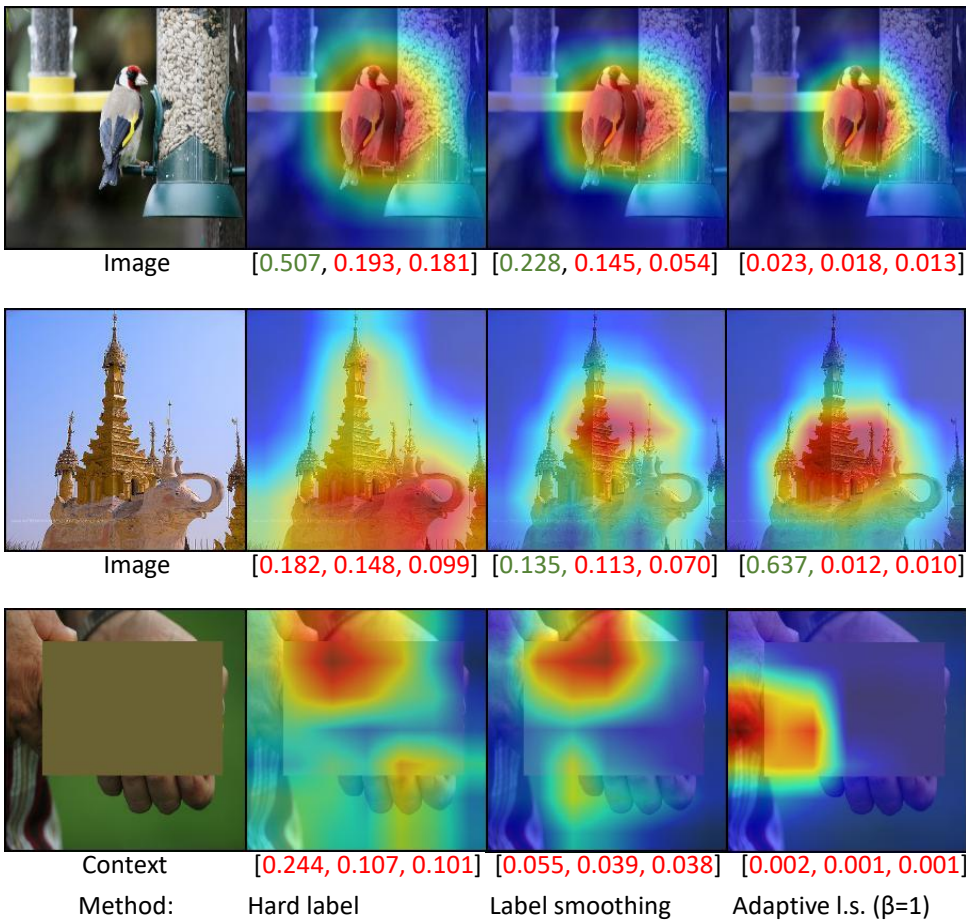
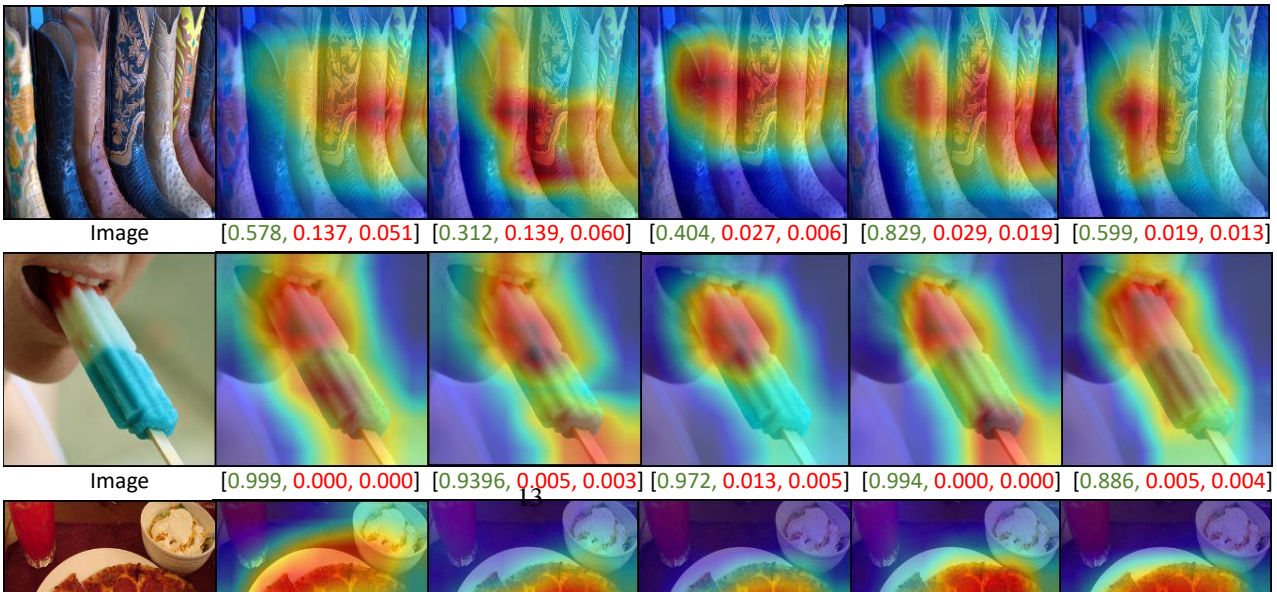


Figure 4: Examples of class activation maps (CAMs). These were obtained using the implementation of (Chattopadhay et al., 2018). Two columns on the left show results for baseline CNNs using hard labels and standard label smoothing. Our approach, *adaptive* label smoothing ("adaptive l.s."), is illustrated in the last column on the right. Our technique produces high-entropy predictions and shows an improved localization performance. The values under each CAM represent the top three probabilities, with green indicating the pertinent class and red indicating an incorrect prediction.

A APPENDIX



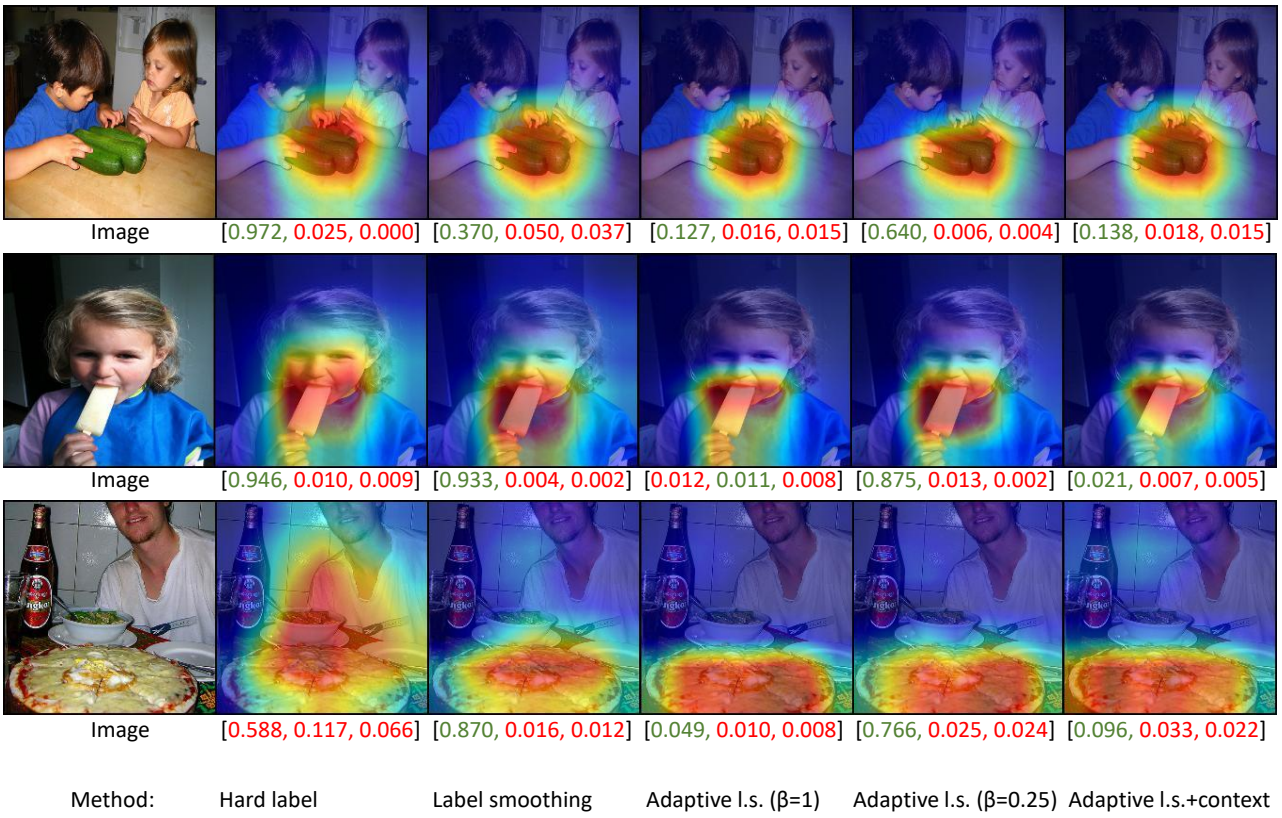


Figure 6: Examples of class activation maps (CAMs). These were obtained using the implementation of Chattopadhay et al. (2018). The second and third columns from the left show results for baseline CNNs using hard labels and standard label smoothing. Our approach, *adaptive* label smoothing ('Adaptive l.s'), is illustrated in the three rightmost columns. Our technique produces high-entropy predictions and shows an improved localization performance. The values under each CAM represent the top three probabilities, with green indicating the pertinent class and red indicating an incorrect prediction.

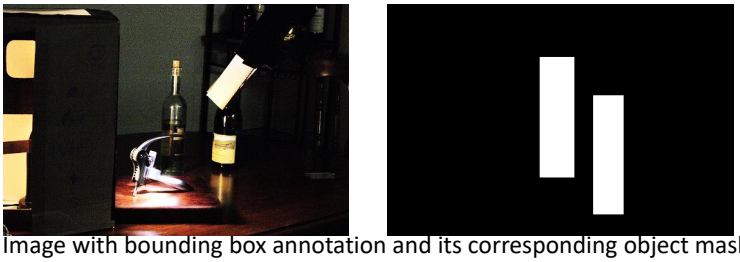
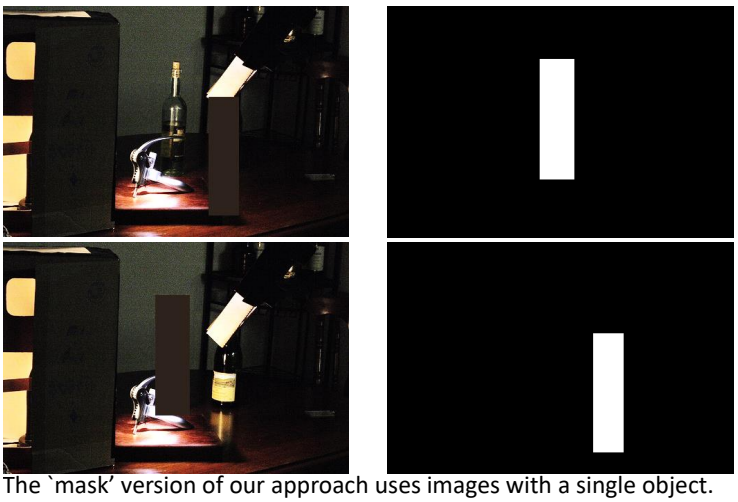
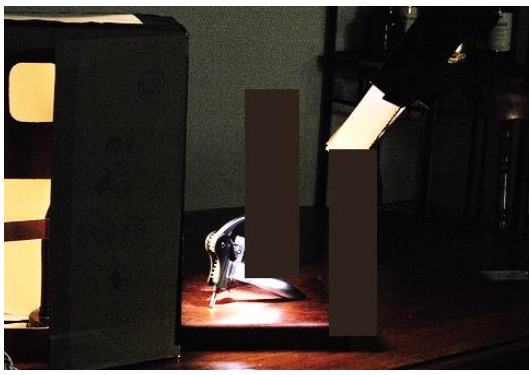


Image with bounding box annotation and its corresponding object mask.



The `mask' version of our approach uses images with a single object.



The `context' version of our approach uses images with all the objects masked out about 15% of the time during training. The label vector for such images (context only) is a vector of uniform distribution.

Figure 7: The first row of images in the left half of the figure are an example of the ImageNet dataset (N=0.474M) that have bounding box annotations. We match the images from the training set of ImageNet-1K dataset with the corresponding '.xml' files included in the ImageNet object detection dataset. We then create object masks for each of the images. When applying any scaling and cropping operation to training samples, we apply the same transformation to the corresponding object masks as well. By counting the number of white pixels, we can determine the object proportion post transformation. We describe the two other approaches in the figure, the 'mask' version of our approach has a single object (for images with multiple bounding box annotations) and this version has 0.528M samples. Our approach helps generate accurate labels during training and penalizes low-entropy (high-confidence) predictions for context-only images like the example on the right half of the figure.