Guided Zoom: Questioning Network Evidence for Fine-Grained Classification

Sarah Adel Bargal[*1]
sbargal@bu.edu

Andrea Zunino[*2]
andrea.zunino@iit.it

Vitali Petsiuk[1]
vpetsiuk@bu.edu

Jianming Zhang[3]
jianmzha@adobe.com

Kate Saenko[1]
saenko@bu.edu

Vittorio Murino[2,4]
vittorio.murino@iit.it

Stan Sclaroff[1]
sclaroff@bu.edu

[1] Department of Computer Science
Boston University

[2] Pattern Analysis & Computer Vision
Istituto Italiano di Tecnologia

[3] Adobe Research

[4] Department of Computer Science
University of Verona

Abstract

We propose Guided Zoom, an approach that utilizes spatial grounding of a model's decision to make more informed predictions. It does so by making sure the model has "the right reasons" for a prediction, defined as reasons that are coherent with those used to make similar correct decisions at training time. The reason/evidence upon which a deep convolutional neural network makes a prediction is defined to be the spatial grounding, in the pixel space, for a specific class conditional probability in the model output. Guided Zoom examines how reasonable such evidence is for each of the top-k predicted classes, rather than solely trusting the top-1 prediction. We show that Guided Zoom improves the classification accuracy of a deep convolutional neural network model and obtains state-of-the-art results on three fine-grained classification benchmark datasets.

1 Introduction

For state-of-the-art deep single-label classification models, the correct class is often in the top-k predictions, leading to a top-k ($k = 2, 3, 4, \dots$) accuracy that is significantly higher than the top-1 accuracy. This accuracy gap is more pronounced in fine-grained classification tasks, where the differences between classes are quite subtle [11, 12, 17, 28]. For example, the Stanford Dogs fine-grained dataset on which we report results has a top-1 accuracy of 86.9% and a top-5 accuracy of 98.9%. Exploiting the information provided in the top-k predicted classes can boost the final prediction of a model.
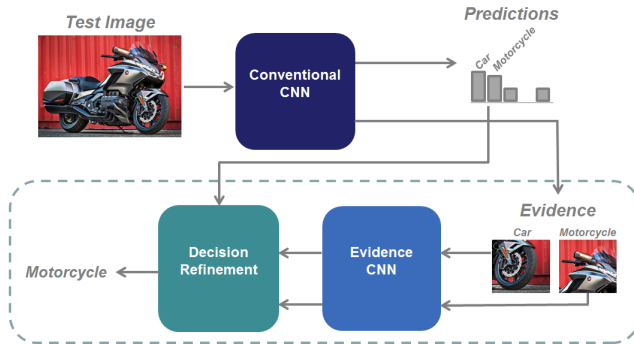
* Equal contribution

Figure 1: Pipeline of `Guided Zoom`. A conventional CNN outputs class conditional probabilities for an input image. Salient patches could reveal that evidence is weak. We refine the class prediction of the conventional CNN by introducing two modules: 1) *Evidence CNN* determines the consistency between the evidence of a test image prediction and that of correctly classified training examples of the same class. 2) *Decision Refinement* uses the output of *Evidence CNN* to refine the prediction of the conventional CNN.

In this work, we do not completely trust a Convolutional Neural Network (CNN) model's top-1 prediction as it does not solely depend on the visual evidence in the input image, but can depend on other artifacts such as dataset bias or unbalanced training data. Instead, we target answering the following question: *is the evidence upon which the prediction is made reasonable?* Evidence is defined to be the grounding, in pixel space, for a specific class conditional probability in the model output. We propose `Guided Zoom`, which uses evidence grounding as the signal to localize discriminative image regions and to assesses how much one can trust a CNN prediction over another.

Since fine-grained classification requires focusing on details, the localization of discriminative object parts is crucial. Supervised approaches utilize part bounding box annotations [9, 30, 32] or have humans in the loop to help reveal discriminative parts [5]. Our approach does not require part annotations, thus it is more scalable compared to supervised approaches. Weakly supervised attention based methods require training CNNs with attention module(s) [7, 18, 24, 36], while our approach is a generic framework able to improve the performance of any CNN model, even if it was not trained using an attention module.

`Guided Zoom` zooms into the evidence used to make a preliminary decision at test time and compares it with a reference pool of (evidence, prediction) pairs. The pool is constructed by accumulating (evidence, prediction) pairs for correctly classified training examples, providing additional free supervision from training data. As demonstrated in Fig. 1, our approach examines whether or not the evidence used to make the prediction is coherent with training evidence of correctly classified images. This is performed by the *Evidence CNN* module, which aids the *Decision Refinement* module to come up with a refined prediction. The main goal of `Guided Zoom` is ensuring that evidence of the refined class prediction is more coherent with the training evidence of that class, compared to evidence of any of the other candidate top classes (see Fig. 2). By examining network evidence, we demonstrate improved accuracy and achieve state-of-the-art results on three fine-grained classification benchmark datasets.
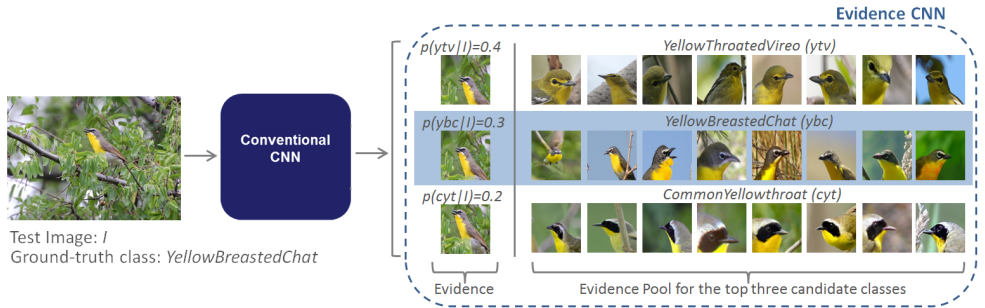
Figure 2: A conventional CNN is used to obtain salient image regions that highlight the evidence for predictions, together with the predicted class conditional probabilities. In our model, fine-grained classification decisions are improved by comparing consistency of the evidence for the incoming test image with the evidence seen for correct classifications in training. In this demonstration, although the conventional CNN predicts with highest probability the class *YellowThroatedVireo*, *Evidence CNN* is able to provide guidance for predicting the ground-truth class *YellowBreastedChat* (highlighted in blue) due to visual similarity of the evidence of this class with that of the generated evidence pool.

2 Related Work

Evidence Grounding. The evidence behind a deep model's prediction for visual data results in highlighting the importance of image regions in making a prediction. Fong *et al.* [6] exhaustively perturb image regions to find the most discriminative evidence. Petsiuk *et al.* [19] probe black-box CNN models with randomly masked instances of an image to find class-specific evidence. While such methods are bottom-up approaches, others are top-down and start from a high-level feature representation and then propagate down to the image level [21, 22, 23, 29, 31, 37]. For example, Selvaraju *et al.* [21] exploit a weighted sum of the last convolutional feature maps to obtain the class activation maps. Zhang *et al.* [31] use network weights and activations to obtain class activation maps, then highlight image cues unique to a specific class by eliminating evidence that also contributes to other classes.

The result of evidence grounding is often referred to as saliency. Saliency is being widely used for many computer vision tasks including spatial semantic segmentation [15, 27, 38], spatial object localization [31, 33], and temporal action localization [1]. Saliency has been less exploited for improving model classification. Cao *et al.* [3] use weakly supervised saliency to feed highly salient regions into the same model that generated them to get more prediction probabilities for the same image and improve classification accuracy at test time. In contrast, we use weakly supervised saliency to examine whether the obtained evidence is coherent with the evidence used at training time for correctly classified examples. Zunino *et al.* [39] use spatial grounding at training time to improve model classification by dropping neurons corresponding to high-saliency patterns for regularization. In contrast, we propose an approach to improve model classification at test time.

Fine-grained classification. The key module in fine-grained classification is finding discriminative parts. Some approaches use supervision to find such discriminative features, *i.e.* use annotations for whole objects and/or for semantic parts. Zhang *et al.* [32] train part models such that the head/body can be compared; however, this requires a lot of annotation

of parts. Krause *et al*. [13] use whole object annotations and no part annotations. Branson *et al*. [2] normalize the pose of object parts before computing a deep representation for them. Zhang *et al*. [30] introduce part-abstraction layers in the deep classification model, enabling weight sharing between the two tasks. Huang *et al*. [9] introduce a part-stacked CNN which encodes part and whole object cues in parallel based on supervised part localization. Wang *et al*. [26] retrieve neighboring images from the dataset, those having similar object pose, and automatically mine discriminative triplets of patches with geometric constraints as the image representation. Deng *et al*. [5] include humans in the loop to help select discriminative features. Subsequent work of Krause *et al*. [14] does not use whole or part annotations, but augments fine-grained datasets by collecting web images and experimenting with filtered and unfiltered versions of them. Wang *et al*. [25] use an ontology tree to obtain hierarchical multi-granularity labels. In contrast to such approaches, we do not require any object or part annotations at train or test time and do not use additional data or hierarchical labels.

Other approaches are weakly supervised. Such approaches only require an image label, and our approach lies in this category. Lin *et al*. [16] demonstrate the applicability of a bilinear CNN model in the fine-grained classification task. Recasens *et al*. [20] propose a distortion layer based on saliency to improve the input image sampling that demonstrates improvement for the fine-grained classification task. Sun *et al*. [24] implement an attention module that learns to localize different parts and a correlation module that enforces coherency in the correlations among different parts in training. Fu *et al*. [7] learn where to focus by recurrently zooming into one location from coarse to fine using a recurrent attention CNN. In contrast, we are able to zoom into multiple image locations. Zhang *et al*. [34] use convolutional filters as part detectors since the responses of distinctive filters usually focus on consistent parts. Zhao *et al*. [35] use a recurrent soft attention mechanism that focuses on different parts of the image at every time step. This work enforces a constraint to minimize the overlap of attention maps used in adjacent time steps to increase the diversity of part selection. Zheng *et al*. [36] implement a multiple attention convolutional neural network with a fully-connected layer, combining the softmax for each part with one classification loss function. Cui *et al*. [4] introduce a kernel pooling scheme and also demonstrate benefit to the fine-grained classification task. Jaderberg *et al*. [10] introduce spatial transformers for convolutional neural networks that learn invariance to translation, scale, rotation and more generic warping, showing improvement for the task of fine-grained classification.

In contrast, our approach assesses whether the network evidence used to make a prediction is reasonable, *i.e*. if it is coherent with the evidence of correctly classified training examples of the same class. We use multiple salient regions, thus eliminating error propagation from incorrect initial saliency localization, and implicitly enforce part-label correlations enabling the model to make more informed predictions at test time.

3 Guided Zoom

Guided Zoom uses the evidence of a prediction to improve classification performance by comparing the coherence of such evidence with a pool of "reasonable" class-specific evidence. We now describe how Guided Zoom utilizes multiple discriminative evidence, does not require part annotations, and implicitly enforces part correlations. This is done through the main modules depicted in Fig. 1: *Evidence CNN* and *Decision Refinement*.

Evidence CNN. Conventional CNNs trained for image classification output class conditional probabilities upon which predictions are made. The class conditional probabilities are

the result of some corresponding evidence in the input image. From correctly classified training examples, we generate a reference pool $\mathcal{P}$ of (evidence, prediction) pairs over which the *Evidence CNN* will be trained for the same classification task. We recover/ground such evidence using a top-down spatial grounding technique: contrastive Excitation Backprop (*c*EB) [31]. Starting with a prior probability distribution, *c*EB passes top-down signals through excitatory connections (having non-negative weights) of a CNN. Recursively propagating the top-down signal layer by layer, *c*EB computes class-specific discriminative saliency maps from any intermediate layer in a partial single backward pass. This is done by setting the prior distribution in correspondence with the correct class to produce a *c*EB saliency map for it as described in the experiments section. We extract one 150x150 pixel patch from the original image around the resulting peak saliency. This patch highlights the most discriminative evidence. However, the next most discriminative patches may also be good additional evidence for differentiating fine-grained categories.

Grounding techniques only highlight part(s) of an object. However, a more inclusive segmentation map can be extracted from the already trained model at test time using adversarial erasing. Inspired by the adversarial erasing work of Wei *et al.* [27] in the domain of object segmentation, we extend its application to the domain of image classification by providing a data augmentation technique for network evidence. We augment our reference pool with patches resulting from performing an iterative adversarial erasing of the most discriminative evidence from the image. We notice that adversarial erasing results in implicit part localization from the most to least discriminative parts. Fig. 3 (right) shows the patches extracted from two iterations of adversarial saliency erasing for sample images belonging to the class *Chihuahua* from the Stanford Dogs Dataset. All patches (parts) extracted from this process inherit the ground-truth label of the original image. By labeling different parts with the same image ground-truth label, we are implicitly forcing part-label correlations in *Evidence CNN*.

Including such additional evidence in our reference pool gives a richer description of the examined classes compared to models that recursively zoom into one location while ignoring other discriminative cues [7]. We note that we add an evidence patch to the reference pool only if the removal of the previous salient patch does not affect the correct classification of the sample image. Erasing is performed by adding a black-filled 85x85 pixel square on the previous most salient evidence to encourage a highlight of the next salient evidence. This process is depicted in Fig. 3 (left) for a sample bird species, dog species, and aircraft model.

Assuming n training samples, for each sample s^i where $i \in 1, \ldots, n$ we have $l + 1$ evidence patches in the reference pool $e_0^i, \ldots, e_l^i$. Patch e_0^i is the most discriminative initial evidence, and $e_1^i, \ldots, e_l^i$ is the set of the l next discriminative evidence patches where $l \leq L$ and L is the number of adversarial erasing iterations performed ($L = 2$ is used in our experiments). For example, e_2^i is the third most-discriminative evidence, after the erasing of e_0^i and e_1^i from the original image. We then train a CNN model, the *Evidence CNN*, on the classification of the generated evidence pool $\mathcal{P}$.

Decision Refinement. At test time, we analyze whether the evidence upon which a prediction is made is reasonable. We do so by examining the consistency of a test (evidence, prediction) with our reference pool that is used to train *Evidence CNN*. We exploit the visual evidence used for each of the top-k predictions for prediction refinement. The refined prediction will be inclined toward each of the top-k classes by an amount proportional to how coherent its evidence is with the reference pool. For example, if the (evidence, prediction) of the second-top predicted class is more coherent with the reference pool of this class, then the refined prediction will be more inclined toward the second-top class.

Assuming test image s^j, where $j \in 1, \ldots, m$ and m is the number of testing examples,
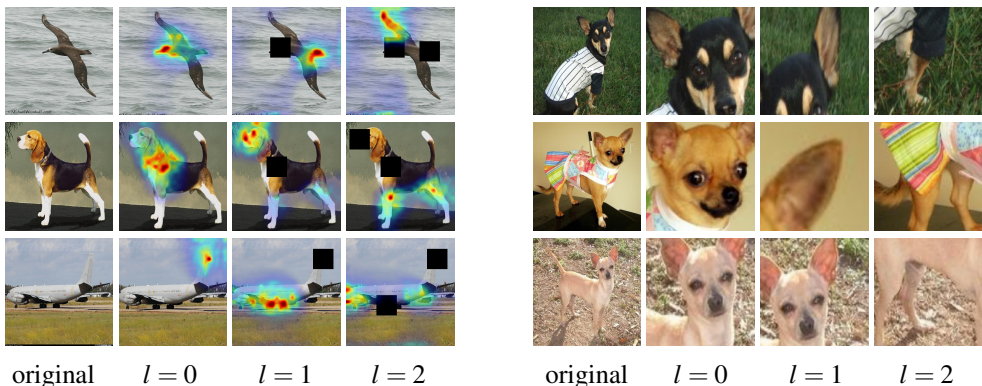
Figure 3: **Adversarial Erasing.** Left: Sample image from each dataset to demonstrate the extraction of patches during two rounds of adversarial erasing: finding the first ($l = 0$), second ($l = 1$), and third ($l = 2$) most-salient evidence. For example, the most salient evidence for the bird image is the head, followed by the tail, followed by the right wing. Right: Implicit part detection obtained as a result of two iterations of adversarial erasing for three images from the dog class *Chihuahua*. Assigning the same class label to the different parts of a single dog image enforces implicit part-label correlation.

s^j is passed through the conventional CNN resulting in $v^{j,0}$, a vector of class conditional probabilities having some top-k classes $c_1, \ldots, c_k$ to be considered for the prediction refinement. We obtain the evidence for each of the top-k predicted classes $e_0^{j,c_1}, \ldots, e_0^{j,c_k}$, and pass each one through the *Evidence CNN* to get the output class conditional probability vectors $v_0^{j,c_1}, \ldots, v_0^{j,c_k}$. We then perform adversarial erasing to get the next most salient evidence $e_l^{j,c_1}, \ldots, e_l^{j,c_k}$ and their corresponding class conditional probability vectors $v_l^{j,c_1}, \ldots, v_l^{j,c_k}$, for $l \in 1, \ldots, L$. Finally, we compute a weighted combination of all class conditional probability vectors proportional to their saliency (a lower l has more discriminative evidence and is therefore assigned a higher weight w_l). The estimated, refined class c_{ref}^j is determined as the class having the maximum aggregate prediction in the weighted combination. Algorithm 1 presents the steps used for decision refinement.

4 Experiments

In this section, we first present the fine-grained benchmark datasets we use to evaluate Guided Zoom. We then present the architecture and setup of our experiments, followed by a discussion of our experimental results. We note that although the datasets provide part annotations, we only use image-level class labels.

Datasets. We report experimental results on three fine-grained classification benchmark datasets following [4, 7, 24, 35, 36].

- CaltechUCSD (CUB-200-2011) Birds Dataset [28] is a fine-grained dataset of 200 bird species consisting of ∼12K annotated images, split into ∼6K training images and ∼6K testing images.

Algorithm 1: Decision Refinement

Input: $s^j, j \in 1, \ldots, m$ testing images, pre-trained conventional CNN, pre-trained *Evidence CNN*, Grounding Method (GM), weights $w, w_0, \ldots, w_L$

Output: Refined class for s^j: c^j_{ref}

Procedure:

1 For every test example $s^j, j \in 1, \ldots, m$

2    $v^{j,0} :=$ conventional CNN(s^j)

3    $tot^j := w * v^{j,0}$

4   For $t \in c_1, \ldots, c_k$, the top-k classes of $v^{j,0}$

5      $e_0^{j,t} :=$ GM(s^j) w.r.t. class t

6      $v_0^{j,t} :=$ *Evidence CNN*$(e_0^{j,t})$

7      $tot^j[t] := tot^j[t] + w_0 * v_0^{j,t}[t]$

8   For $l \in 1, \ldots, L$

9   Adversarially erase $e_{l-1}^{j,t}$ from s^j

10       $e_l^{j,t} :=$ GM(s^j) w.r.t. class t

11       $v_l^{j,t} :=$ *Evidence CNN*$(e_l^{j,t})$

12       $tot^j[t] := tot^j[t] + w_l * v_l^{j,t}[t]$

13   $c^j_{ref} := \underset{c_1:c_k}{\operatorname{argmax}}(tot^j)$

- Stanford Dogs Dataset [[11]] is a fine-grained dataset of 120 dog species. This dataset includes ∼20K annotated images split into ∼12K and ∼8.5K images for training and testing respectively.

- FGVC-Aircraft [[17]] is a fine-grained dataset of 100 different aircraft variants consisting of 10K annotated images, split into ∼7K training images and ∼3K testing images.

Architecture and Setup. To validate the benefit of `Guided Zoom`, we first purposely use a simple CNN baseline with a vanilla training scheme. We use a ResNet-101 [[8]] network as the conventional CNN and baseline, extending the input size from the default 224x224 pixel to 448x448 pixel following [[7, 14, 24]]. The 448x448 pixel input image is a random crop from a 475x475 pixel input image at training time, and a center crop from a 475x475 pixel input image at test time. The reference pool is generated by extracting evidence patches from training images using a cEB saliency map computed at the *res4a*. We also then use the more complex MA-CNN architecture [[36]] to demonstrate the generic nature of `Guided Zoom`. This follows the same input image sizes used for ResNet-101, and compute cEB saliency maps at the *conv3_1* layer.

For the *Evidence CNN*, we use a ResNet-101 architecture, but use the standard 224x224

Method	Part / Whole Annotation	Multiple Attention	Top-1 Accuracy (%)
DVAN [35]	x	✓	79.0
PA-CNN [13]	✓	✓	82.8
MG-CNN [25]	✓	✓	83.0
B-CNN [16]	x	x	84.1
RA-CNN [7]	x	x	85.3
PN-CNN [2]	✓	✓	85.4
OSME + MAMC [24]	x	✓	86.5
MA-CNN [36]	x	✓	86.5
ResNet-101	x	x	82.3
Guided Zoom (ResNet-101, k=3)	x	✓	85.0
Guided Zoom (ResNet-101, k=5)	x	✓	85.4
Guided Zoom (MA-CNN, k=3)	x	✓	87.6
Guided Zoom (MA-CNN, k=5)	x	✓	**87.7**

Table 1: **CUB-200-2011 Birds Dataset.** We compare our classification accuracy with state-of-the-art weakly-supervised methods (do not use any sort of annotation apart from the image label) and some representative methods that use additional supervision such as part annotations for fine-grained classification of this dataset. We indicate which methods use multiple parts, and which focus on a single part using the multiple attention flag. Using part annotations implicitly entails multiple attention. We present results for our approach for k=3,5 using the top 3 (or 5) candidate classes to refine the final prediction.

pixel input size to keep the patches close to their original image resolution. This is a random crop from a 256x256 pixel input image at training time, and a center crop from a 256x256 pixel input image at test time. For both the conventional and *Evidence* CNNs, and for all the three datasets, we use stochastic gradient descent, a batch size of 64, a starting learning rate of 0.001, multiplied by 0.1 every 10K iterations for 30K iterations, and momentum of 0.9.

We demonstrate the benefit of using evidence information from the top-3 and top-5 predicted classes, so we set $k = 3, 5$ in our experiments. We perform two rounds of adversarial erasing in testing; we set $L = 2$, and $w = 0.4, w_0 = 0.3, w_1 = 0.2$, and $w_2 = 0.1$ giving lower weights to less discriminative evidence.

Results. In this section, we demonstrate how training our *Evidence CNN* module benefits from implicit part detection, a direct consequence of adversarial erasing that extracts the next most-salient evidence. Our framework targets providing complementary zooming on salient parts and use them for decision refinement.

Table 1 presents the results for the CUB-200-2011 Birds dataset. Utilizing the top-3 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 82.3% to 85.0%. Utilizing the top-5 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 82.3% to 85.4%. Being a generic framework, we implement Guided Zoom to the image stream of the multi-zoom approach MA-CNN, obtaining state-of-the-art results on the CUB-200-2011 dataset.

Table 2 presents the results for the Stanford Dogs dataset. Utilizing the top-3 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 86.9% to 88.4%. Utilizing the top-5 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 86.9% to 88.5%. We achieve state-of-the-art results on the Stanford Dogs dataset.

Method	Part / Whole Annotation	Multiple Attention	Top-1 Accuracy (%)
DVAN [35]	x	✓	81.5
OSME + MAMC [24]	x	✓	85.2
RA-CNN [7]	x	x	87.3
Ours ResNet-101	x	x	86.9
Guided Zoom (ResNet-101, k=3)	x	✓	88.4
Guided Zoom (ResNet-101, k=5)	x	✓	**88.5**

Table 2: **Stanford Dogs Dataset.** We compare our classification accuracy with state-of-the-art weakly-supervised methods (do not use any sort of annotation apart from the image label). We indicate which methods use multiple parts, and which focus on a single part using the multiple attention flag. Using part annotations implicitly entails multiple attention. We present results for our approach for k=3,5 using the top 3 (or 5) candidate classes to refine the final prediction.

Method	Part / Whole Annotation	Multiple Attention	Top-1 Accuracy (%)
B-CNN [16]	x	x	84.1
MG-CNN [25]	✓	✓	86.6
RA-CNN [7]	x	x	88.2
MDTP [26]	✓	✓	88.4
MA-CNN [36]	x	✓	89.9
Ours ResNet-101	x	x	87.5
Guided Zoom (ResNet-101, k=3)	x	✓	89.1
Guided Zoom (ResNet-101, k=5)	x	✓	89.1
Guided Zoom (MA-CNN, k=3)	x	✓	**90.7**
Guided Zoom (MA-CNN, k=5)	x	✓	**90.7**

Table 3: **FGVC-Aircraft Dataset.** We compare our classification accuracy with state-of-the-art weakly-supervised methods (do not use any sort of annotation apart from the image label) and some representative methods that use additional supervision such as part annotations for fine-grained classification of this dataset. We indicate which methods use multiple parts, and which focus on a single part using the multiple attention flag. Using part annotations implicitly entails multiple attention. We present results for our approach for k=3,5 using the top 3 (or 5) candidate classes to refine the final prediction.

Table 3 presents the results for the FGVC-Aircraft dataset. Utilizing the top-3 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 87.5% to 89.1%. Utilizing the top-5 class predictions together with their associated evidence, Guided Zoom boosts the top-1 accuracy from 87.5% to 89.0%. Guided Zoom boosts the top-1 classification accuracy of MA-CNN obtaining state-of-the-art results on the FGVC-Aircraft dataset.

Guided Zoom outperforms RA-CNN on all three datasets. From this we can conclude that our multi-zooming is more beneficial than a single recursive zoom. In addition, we demonstrate that Guided Zoom further improves the performance of the multi-zoom approach MA-CNN. Being a generic framework, Guided Zoom could be applied to any deep convolutional model for decision refinement within the top-k predictions.

Conclusion

In this work, we propose `Guided Zoom`, which utilizes explicit spatial grounding to refine a model's prediction at test time. Our refinement module selects one of the top-k model predictions having the most reasonable (evidence, prediction) pair, where "most reasonable" is defined as the most consistent with respect to a pre-defined pool of training (evidence, prediction) pairs. The pool is populated by iteratively grounding and adversarially erasing the evidence for a correct prediction made by a conventional CNN. `Guided Zoom` achieves state-of-the-art results on three fine-grained classification benchmark datasets.

Acknowledgments

References

[1] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for RNNs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014.

[3] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[6] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2018.

[7] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[9] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proc. Advances in Neural Information Processing Systems*, 2015.

[11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRw)*, 2011.

[12] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRw)*, 2013.

[13] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proc. European Conf. on Computer Vision (ECCV)*, 2016.

[15] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2015.

[17] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[18] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Proc. Advances in Neural Information Processing Systems*, 2014.

[19] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proc. British Machine Vision Conference (BMVC)*, 2018.

[20] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2017.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[24] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[25] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2015.

[26] Yaming Wang, Jonghyun Choi, Vlad Morariu, and Larry S Davis. Mining discriminative triplets of patches for fine-grained classification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. In *Technical Report CNS-TR-2010-001, California Institute of Technology*, 2010.

[29] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. European Conf. on Computer Vision (ECCV)*, 2014.

[30] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision (IJCV)*, 126(10):1084–1102, 2018.

[32] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *Proc. European Conf. on Computer Vision (ECCV)*, 2014.

[33] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. on Multimedia*, 19(6):1245–1256, 2017.

[36] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2017.

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[38] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Andrea Zunino, Sarah Adel Bargal, Pietro Morerio, Jianming Zhang, Stan Sclaroff, and Vittorio Murino. Excitation dropout: Encouraging plasticity in deep neural networks. *arXiv preprint arXiv:1805.09092*, 2018.