Are CNN Predictions based on Reasonable Evidence?

Sarah Adel Bargal*[1], Andrea Zunino*[2], Vitali Petsiuk[1], Jianming Zhang[3],
Kate Saenko[1], Vittorio Murino[2,4], Stan Sclaroff[1]

[1]Department of Computer Science, Boston University [2]Pattern Analysis & Computer Vision (PAVIS),
Istituto Italiano di Tecnologia [3]Adobe Research [4]Computer Science Department, Università di Verona
{sbargal,vpetsiuk,saenko,sclaroff}@bu.edu, {andrea.zunino,vittorio.murino}@iit.it, jianmzha@adobe.com

We propose `Guided Zoom`, an approach that utilizes spatial grounding to make more informed predictions. It does so by making sure the model has "the right reasons" for a prediction, being defined as reasons that are coherent with those used to make similar correct decisions at training time. The reason/evidence upon which a deep 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` questions how reasonable the evidence used to make a prediction is. We show that `Guided Zoom` results in the refinement of a model's classification accuracy on two fine-grained classification 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 is also more crucial in fine-grained classification tasks, where the differences between classes are quite subtle. 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. In this work, we do not completely trust the 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 exploit the discriminative visual evidence used for each of the top-k predictions for decision refinement.

Examples of fine-grained classes present in the literature are breeds of animals [7], birds [14], models of aircraft [10] and vehicles [9]. Since fine-grained classification requires focusing on details, the localization of salient parts is crucial. This has been addressed using supervised approaches that utilize part bounding box annotations [15, 17, 5] or have humans in the loop to help reveal discriminative parts [3]. Part localization has also been addressed using weakly supervised approaches [4, 12, 19, 6], solely relying on image labels during both training and testing. Another class of works attend to a recursively zoomed location [4, 11], while other methods use multiple attention mechanisms [12, 19]. Some approaches enforce correlations between parts [12, 6], while others do not consider this possible source of information [8, 4].

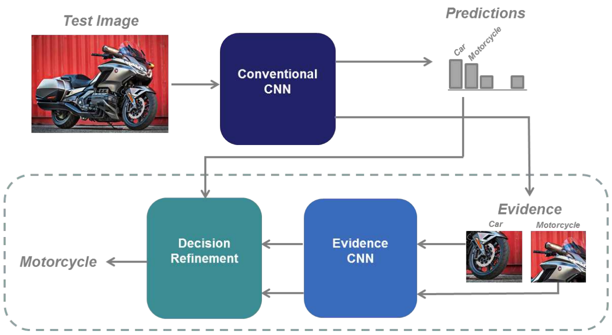

Figure 1: Pipeline of `Guided Zoom`. A conventional CNN outputs class conditional probabilities for an input image. Salient patches could reveal that evidence is weak. We refine the prediction of the conventional CNN by introducing two modules: 1) *Evidence CNN* determines the consistency between the 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 want to answer 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. The evidence proposed here is in the form of a saliency map resulting from weak supervision. It

*Equal contribution

is directly obtained using grounding approaches that utilize a network's internal representation and a dataset's image-level annotation.

Saliency is widely used for many computer vision tasks including spatial semantic segmentation, spatial object localization, and temporal action localization. However, saliency has been less exploited for improving model classification. Cao *et al*. [1] use weakly supervised saliency to feedback 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 evidence grounding as the signal to a module that assesses how much one can trust a Convolutional Neural Network (CNN) prediction over another.

We propose `Guided Zoom`, an approach that utilizes spatial grounding to refine model predictions in fine-grained classification scenarios. `Guided Zoom` zooms in on the evidence used to make a preliminary decision at test time and compares it with the evidence of correct predictions made at training time. As demonstrated in Fig. 1, we propose not to solely rely on the prediction a conventional CNN produces, but to examine 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 desired goal in `Guided Zoom` is that the evidence of the refined class prediction is more coherent with the training evidence of that class, than the evidence of any of the other candidate top classes.

Our approach does not require part annotations, thus it is more scalable compared to supervised approaches. Moreover, our approach uses multiple salient regions and therefore does not propagate errors from an incorrect initial saliency localization, while implicitly enforcing part correlations enabling models to make more informed predictions.

As the experiments of Wei *et al*. [13] suggest, although only part(s) of an object will be highlighted in the evidence, a more inclusive segmentation map can be extracted from the already trained model at test time. We follow their strategy of adversarial erasing to obtain a rich representation for the *Evidence CNN* module. By questioning network evidence, we demonstrate refined accuracy on two fine-grained classification benchmark datasets.

2. Guided Zoom

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. We recover/ground such evidence using the spatial grounding method contrastive Excitation Backprop (*c*EB) [16]. 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.

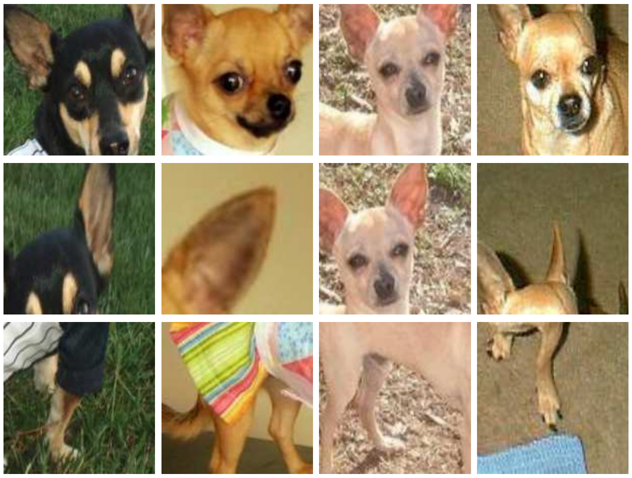

Figure 2: Implicit part detection obtained as a result of two iterations of adversarial erasing. The first row shows the most salient patches of four images from the class *Chihuahua* in the Stanford Dogs dataset. The second row shows the second most salient patches, and the third row shows the third most salient patches for the same four images. Assigning the same class label to the different parts of a single dog image enforces implicit part-label correlation.

We generate a reference pool, $\mathcal{P}$ of (evidence, prediction) pairs over which *Evidence CNN* will be trained for the same classification task. Pairs in the pool $\mathcal{P}$ are extracted for correctly classified training examples using the grounding method *c*EB. This is done by setting the prior distribution in correspondence with the correct class to produce a *c*EB saliency map for it. We extract 150x150-pixel patches from the original image around the resulting peak saliency. For example, the most discriminative evidence to differentiate dogs tends to be the face. However, the next most discriminative patches may also be good additional evidence for differentiating fine-grained categories.

Inspired by the adversarial erasing work of Wei *et al*. [13], 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. 2 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

	Method	Part / Whole Annotation	Multiple Attention	CUB-200-2011 Dataset Top-1 Accuracy (%)	Stanford Dogs Dataset Top-1 Accuracy (%)
	RA-CNN [4]	x	x	85.3	87.3
	OSME + MAMC [12]	x	✓	**86.5**	85.2
	MA-CNN [19]	x	✓	**86.5**	-
Ours	ResNet-101 Baseline	x	x	82.3	86.9
	Guided Zoom (k=3)	x	✓	85.0	88.4
	Guided Zoom (k=5)	x	✓	85.4	**88.5**

Table 1: We present results for our approach for k=3,5; using the top 3 (or 5) candidate classes to refine the final prediction. Guided Zoom improves the ResNet-101 Baseline by at least 2.7% for the CUB-200-2011 dataset and at least 1.5% for the Stanford Dogs dataset. We also 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.

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 and ignore the less discriminative cues [4]. We note that we add an evidence patch to the reference pool only if the removal of previous salient patch does not affect the correct classification of the sample s^i. Erasing is performed by adding a black-filled 85x85-pixel square on the previous most salient evidence to encourage a highlight of the next most salient evidence.

Assuming n training samples, for each sample s^i where $i \in 1, \dots, n$ we have $l + 1$ evidence patches in the reference pool $e_0^i, \dots, e_l^i$. e_0^i is the most discriminative initial evidence, and $e_1^i, \dots, e_l^i$ is the set of l next discriminative evidence 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, *Evidence CNN*, on 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*. The refined prediction will be biased 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 biased toward the second-top class.

Assuming test image s^j, where $j \in 1, \dots, m$ and m is the number of testing examples, s^j is passed through the conventional CNN resulting in $v^{j,0}$, a vector of class conditional probabilities having some top-k classes $c_1, \dots, c_k$ to be considered for the prediction refinement. We obtain the evidence for each of the top-k predicted classes $e_0^{j,c_1}, \dots, e_0^{j,c_k}$, and pass each one through the *Evidence CNN* to get the following output class conditional probability vectors $v_0^{j,c_1}, \dots, v_0^{j,c_k}$. We then perform adversarial erasing to get the next most salient evidence $e_l^{j,c_1}, \dots, e_l^{j,c_k}$ and their corresponding class conditional probability vectors $v_l^{j,c_1}, \dots, v_l^{j,c_k}$, for $l \in 1, \dots, L$. Finally, we compute a weighted combination of the class conditional probability vectors proportional to their saliency. The estimated, refined class c_{ref}^j is determined as the class having the maximum aggregate prediction in the weighted combination.

3. Experiments

Datasets. We report experimental results on two fine-grained classification datasets following [12, 4, 18, 2, 19]. CaltechUCSD (CUB-200-2011) Birds Dataset [14] is a fine-grained dataset of 200 bird species (∼12K images). Stanford Dogs Dataset [7] is a fine-grained dataset of 120 dog species (∼20K images).

Architecture and Setup. To validate the benefit of Guided Zoom, we purposely use a simple CNN baseline with a vanilla training scheme. We use a ResNet-101 network as the conventional CNN and baseline, extending the input size from the default 224x224-pixel to 448x448-pixel following [12, 4, 8].

For the *Evidence CNN*, we use a ResNet-101 architecture, but use the standard 224x224-pixel input size to keep the patches close to their original image resolution. 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; setting $L = 2$.

Results. We now present results on the fine-grained datasets: CUB-200-2011 Birds and Stanford Dogs. In this section, we demonstrate how training our *Evidence CNN* benefits from using implicit part detection by adversarial erasing to obtain the next most-salient evidence which targets providing complementary zooming on salient parts.

Table 1 presents the results. For the CUB-200-2011 Birds dataset, our conventional CNN (ResNet-101 baseline) achieves 82.3% top-1 accuracy. Utilizing the top-3 (top-5) class predictions together with their associated evidence, `Guided Zoom` boosts the top-1 class accuracy from 82.3% to 85.0% (85.4%). For the Stanford Dogs dataset, our conventional CNN (ResNet-101 baseline) achieves 86.9% top-1 accuracy. Utilizing the top-3 (top-5) class predictions together with their associated evidence, `Guided Zoom` boosts the top-1 accuracy from 86.9% to 88.4% (88.5%), which is state-of-the-art result.

`Guided Zoom` outperforms RA-CNN on both datasets. From this we can conclude that our multi-zooming is more beneficial than a single recursive zoom. `Guided Zoom` outperforms OSME + MAMC on the Stanford Dogs Dataset, but the opposite is true for the CUB-200-2011 Birds Dataset. Being a generic framework, we plan to next apply `Guided Zoom` to further boost performance of state-of-the-art methods.

Conclusion

In this work, we devise a methodology that utilizes explicit spatial grounding to refine a model's prediction at test time. Our refinement module selects one of the top-k model predictions based on which has the most reasonable (evidence, prediction) pair; defined as the most consistent with respect to a pre-defined pool generated once using adversarial erasing of a grounding technique. We find that `Guided Zoom` improves a base model's prediction accuracy.

References

[1] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. 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. 2

[2] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie. Kernel pooling for convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[3] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2013. 1

[4] J. Fu, H. Zheng, and T. 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. 1, 3

[5] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015. 1

[7] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRw)*, 2011. 1, 3

[8] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Proc. European Conf. on Computer Vision (ECCV)*, 2016. 1, 3

[9] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRw)*, 2013. 1

[10] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. In *Technical Report*. 1

[11] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014. 1

[12] M. Sun, Y. Yuan, F. Zhou, and E. Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proc. European Conf. on Computer Vision (ECCV)*, 2018. 1, 3

[13] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. 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. 2

[14] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. In *Technical Report CNS-TR-2010-001, California Institute of Technology*, 2010. 1, 3

[15] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. 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. 1

[16] J. Zhang, S. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision (IJCV)*, 126(10):1084–1102, 2018. 2

[17] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *Proc. European Conf. on Computer Vision (ECCV)*, 2014. 1

[18] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. on Multimedia*, 19(6):1245–1256, 2017. 3

[19] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2017. 1, 3