Impact of Pretrained Networks For Snake Species Classification

Moorthy Gokula Krishnan

Eloop Mobility Solutions, India
gokul@eloop.ai

Abstract. A robust snake species classifier could aid in the treatment of snake bites. In this report, the technique of transfer learning is revisited to understand the significance of the underlying pre-trained network and the supervised datasets used for pre-training. In low data regime, the methodology of transfer learning has been instrumental in building reliable image classifiers. Comparisons are made between the pre-trained networks trained on datasets of different sizes and classes. Performance improves significantly when the pre-trained network is trained on a much larger supervised dataset. Using country metadata improves the performance considerably. In SnakeCLEF2020 challenge, an F1-score of 0.625 was achieved.

Keywords: Snake Species Classification, Computer Vision, Transfer Learning, Convolutional Neural Networks

1 Introduction

Snakebite is the second most deadly neglected tropical disease [1], being responsible for a dramatic humanitarian crisis in global health. Snakebite envenoming (SBE) affects as many as 2.7 million people every year[3], most of whom live in some of the world's remote, poorly developed, and politically marginalized tropical communities. With annual mortality of 81,000 to 138,000 and 400,000 surviving victims with permanent physical and psychological disabilities, SBE is a disease in urgent need of attention. Antivenoms can be life-saving when correctly administered but this often depends on the correct taxonomic identification (i.e. family, genus, and species) of the biting snake. But, snakes are never identified in nearly 50% of cases globally[4]. An automated system that suggests an identification to the healthcare provider from a low-quality photo can speed up the process of treatment. The participants of SnakeCLEF2020 [10] were challenged to build an accurate snake species classifier that works under diverse conditions.

2 Dataset

With the goal of developing biodiversity monitoring systems, LifeCLEF [6] evaluation campaign aims at benchmarking the progress every year in the identification of plants and animals. SnakeCLEF challenge was introduced in 2020 to benchmark the progress in building a snake species classifier. In this challenge, 245,185 training images are provided split into 783 species. As shown in Figure 1, several aspects of snake morphology make this task challenging for computer vision. Evaluation is done using F1-score which ensures the need for better precision and recall over all the species. The trained model is used to infer labels on the test images that are hidden to participants on platform AICrowd [1] directly. The dataset is extremely imbalanced as indicated in Figure 2 with the minimum number of images per class being 17 and the highest class containing 12,201 images. Additional geographical metadata (country and continent) for the image is also provided. All ablation studies were done locally with the given validation set comprising of 14,029 images.

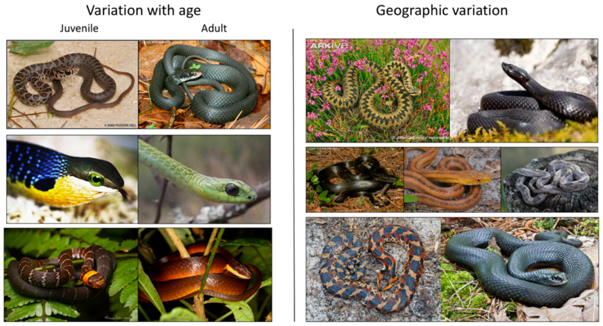

Fig. 1. Snakes diversity [2]

3 Related Work

With the renaissance of deep learning for building image classifiers since 2012 [8], deep convolutional neural networks have become the standard for developing state of the art of image classifiers that work well under diverse conditions given that a large supervised dataset is available. In certain domain-specific cases, the availability of such large scale dataset comprising of millions of images might not be possible. The images might not be readily available, geographically constrained, or rare. In such cases, the technique of transfer learning is used. In this methodology, the network is trained on a different data distribution containing

[1] https://www.aicrowd.com/challenges/snake-species-identification-challenge

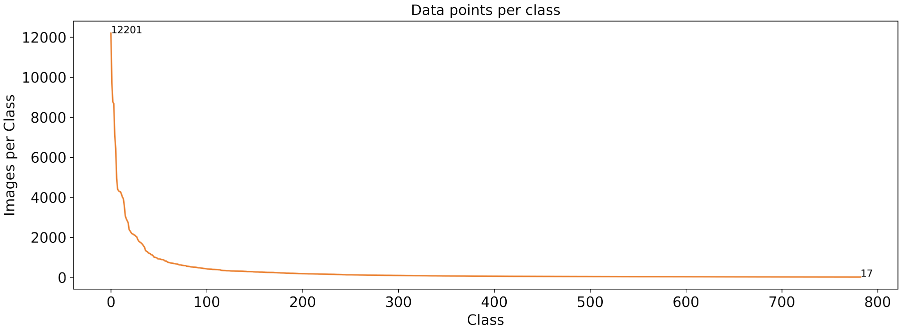

Fig. 2. Data distribution

millions of images and later fine-tuned to domain-specific tasks such as snake species. The dataset used in The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [12] which comprises of 1.4 million images categorized into 1000 classes (ImageNet-1k) is often used to benchmark the performance of image classifiers and the availability of pre-trained models motivates the computer vision community to use the learned representations from the ImageNet-1k dataset. However, ImageNet-1k is a small subset of a much larger dataset containing 14.2 million images categorized into 21,841 classes (ImageNet-21k).

Recently, an extensive study has been performed [7] to understand the impact of these learned representations on the downstream task (fine-tuning to domain-specific image classifier). Corollary to popular belief, larger models trained on larger datasets do not always perform better on the downstream task. The size of the domain-specific dataset plays a crucial role in determining the training strategy and the size of the model. In the context of the SnakeCLEF2020 challenge, experiments were carried out to understand the differences between the models trained on both these datasets (ImageNet-1k and ImageNet-21k).

4 Implementation Details

4.1 Pretrained Classifiers

Vanilla ResNet50-v2 [5] classifier is used for experimentation. Open-source models that were trained on both ImageNet-1k and ImageNet-21k were used. Both of the pre-trained classifiers were trained under the same conditions. Specifically, this involved keeping the hyperparameters, image resolution, and augmentations constant. The fully connected (FC) layer differs depending on the labels specific to the dataset. While fine-tuning, the FC layer is replaced with a domain-specific FC layer randomly initialized.

The following strategies were adopted during training:

- Trained for 10,000 steps.
- The batch size of each step was 512.
- Mixup augmentation was used.
- Staircase based Learning rate scheduler.
- Optimizer: Schocastic gradient descent with momentum 0.9.
- Cross Entropy Loss.

4.2 Training Techniques

Preprocessing The given images are of varied sizes. During the training process, the images are first resized to 512x512x3 dimensions using bilinear interpolation method and a random crop of 456x456x3 was taken. The images were also horizontally flipped with a probability of 0.5. During the validation and the testing process, the images were only resized to 456x456x3 dimensions using bilinear interpolation. The images were also normalized with a standard deviation and mean of 0.5 and 0.5 respectively for training, validation, and testing process.

Batch Accumulation Training can be very inefficient if the mini-batch size is small due to noisy gradients. To accommodate large mini-batch size into GPU memory, batch accumulation is generally used. Gradients are accumulated over 16 steps without updating the model and then updated. Although the size of each mini-batch is 32, the effective mini-batch size is 512.

Learning Rate Learning rate is the crucial hyperparameter to the task of fine-tuning. After linear warmup, a stair-case based learning rate scheduler was used following the hyperrule provided by [7]. Base learning rate (lr_b) of 0.03 was used.

- Step 0-500: Linear warmup : $lr_b * \frac{i}{500}$ where $i = 0, 1, 2..., 500$
- Step 500-10000: lr_b decayed by a factor of 10 at 3000,6000 and 9000 steps

Normalization Group normalization [14] technique along with weight standardization [11] was used. The accuracy of Group normalization is stable across a wide range of batch sizes. It is worth noting that other common techniques like weight decay and dropout were not used.

Augmentation Deep neural networks are prone to undesirable behaviors such as memorization, sensitivity to adversarial examples, and sampling bias. To combat the issues of overfitting, mixup augmentation [15] was used. Mixup strategy trains the network on convex combinations of pairs of examples and their labels. The combination can be controlled by a factor, α. With $\alpha = 0.1$, a sample training image is shown in Figure 3. This favours the network to discriminate between various classes better.

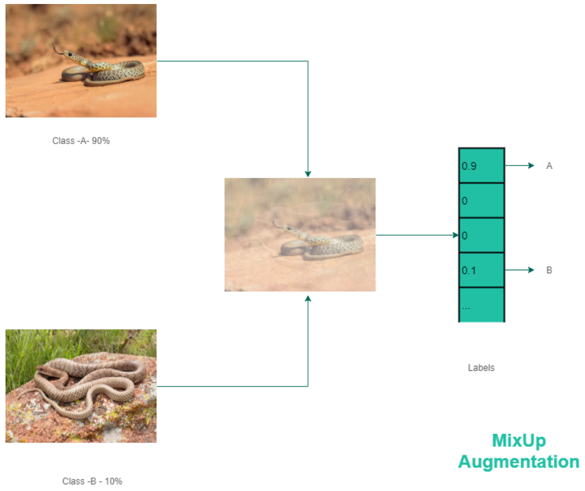

Fig. 3. Augmentation Technique

5 Experiments and Analysis

Fine-tuning is done on two open source models of ResNet50-v2 architecture pre-trained on ImageNet-1k (Model-A) and ImageNet-21k (Model-B) datasets respectively under same training conditions. Results are summarized below:

	Pretrained On	
Metric	**ImageNet-1k**	**ImageNet-21k**
Top-1 Accuracy	68.48 %	79.57 % **+11.09%**
Top-5 Accuracy	87.70%	93.58 % **+5.88%**
F1 score	0.27	0.5813 **+0.3113**

Table 1. Performance comparisons of pre-trained networks

The network pre-trained on ImageNet-21k significantly outperforms its ImageNet-1k counterpart. Especially, classes with fewer data points are discriminated better reflected by the significant improvement in the F1-score.

Model-A disagrees with Model-B for 1,989 images, where Model-B is correct. Also, Model-B disagrees with Model-A for 433 images, where Model-A is correct. By analyzing images with the highest discrepancies (i.e) the images for which Model-B is correct and the probability of the correct species inferred from Model-A is very small, further insights could be gained. An attribution technique [13] to understand which pixels(features) are considered important by the model was performed. The generated saliency maps for top 3 images where models disagree the maximum was chosen. A single gradient step with respect to the target class for the given image was calculated. As shown in Figure 4, by ranking the pixels with respect to the gradients, the saliency maps generated from Model-B tend to be much more concentrated in the area of interest indicating better generalization.

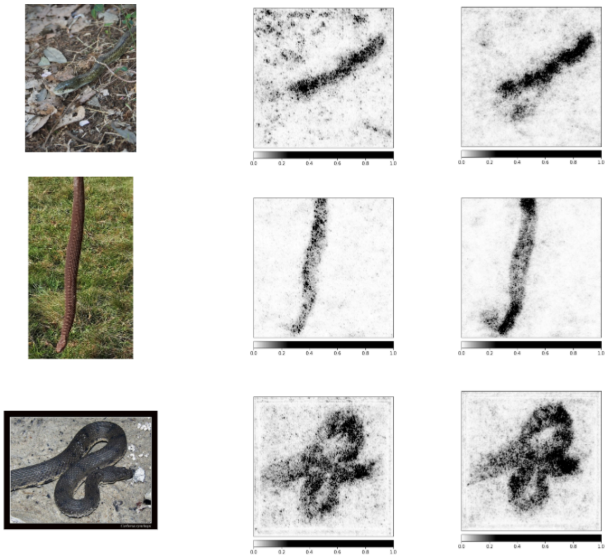

Fig. 4. Generated saliency maps. The first column denotes the images given, second column and third column denote the saliency maps from Model-A and Model-B respectively.

6 Metadata Usage

Several snake species are constrained by their geographical location. Metadata about where the image was taken was given in the form of Country and Continent. The distribution of images per country follows a long tail distribution and is concentrated mostly in the United States Of America (61.42%). Images were taken from 187 countries. In the absence of such information, "UNKNOWN" is marked. The probability of a species given a country is precomputed from the training dataset distribution as follows:

$$p(s \mid c) = t_{sc}/t_c \qquad (1)$$

where:

$p(s \mid c)$ = probability of a species, given a country.
t_{sc} = total number of images belonging to the species, s found in the country, c.
t_c = total number of images found in the country, c.

The generated, $p(s \mid c)$, is provided at [1] along with the source code used for training. It is worth noting that the images marked with "UNKNOWN" data is considered as a country for the purposes of pre-computation. The final probabilities are then adjusted as follows:

$$p_s = p_{\phi_s} * p(s \mid c) \tag{2}$$

where:

p_s = probability of the species, given the image and country.
p_{ϕ_s} = probability of the species inferred from the model.

The probabilites are normalized to ensure a sum of 1. Using this technique on Model-B, the scores improved from 0.5813 to 0.6019.

The test dataset on which the final scores were calculated, follows a data distribution similar to validation dataset and the model achieves an F1-score of 0.625 when tested on AICrowd platform.

7 Conclusion and Future Work

Although the models were trained with the same hyper-parameters, Model-B performs better than Model-A. These results signify the importance of having generalized visual representations before fine-tuning is done on a domain-specific dataset. Label smoothing[9] could improve the performance and can handle the noisy images found in the dataset. Bigger models and stronger augmentations such as rotation and jittering could make the model more resilient.

8 Acknowledgements

The compute resources required to perform the experiments was provided by hostkey.com [2]

[1] https://github.com/GokulEpiphany/snakes-round-4-train/tree/master/metadata

[2] https://www.hostkey.com/

References

1. First medical decision support tool for snake identification based on artificial intelligence and remote collaborative expertise. URL: `https://www.unige.ch/medecine/isg/en/research/one-health/snapp-first-medical-decisionsupport-tool-for-snake-identification-based-on-artificial-intelligence-and-remote-collaborative-expe/`.
2. Image shared on the challenge page. URL: `https://crowdai-shared.s3.eu-central-1.amazonaws.com/markdown_editor/f4e927cb3680ceb410ff825a8c0a53c4_picture2_challenge_datasets.png`.
3. Snakebite envenoming. URL: `https://www.who.int/news-room/fact-sheets/detail/snakebite-envenoming`.
4. Isabelle Bolon, Andrew M. Durso, Sara Botero Mesa, Nicolas Ray, Gabriel Alcoba, François Chappuis, and Rafael Ruiz de Castañeda. Identifying the snake: First scoping review on practices of communities and healthcare providers confronted with snakebite across the world. *PLOS ONE*, 15:1–24, 03 2020. URL: `https://doi.org/10.1371/journal.pone.0229989`.
5. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing.
6. Alexis Joly, Hervé Goëau, Stefan Kahl, Benjamin Deneu, Maximilien Servajean, Elijah Cole, Lukáš Picek, Rafael Ruiz De Castañeda, Isabelle Bolon, Titouan Lorieul, Christophe Botella, Hervé Glotin, Julien Champ, Willem-Pier Vellinga, Fabian-Robert Stöter, Andrew Dorso, Pierre Bonnet, Ivan Eggel, and Henning Müller. Overview of lifeclef 2020: a system-oriented evaluation of automated species identification and species distribution prediction. In *Proceedings of CLEF 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece.*, 2020.
7. Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2019. `arXiv:1912.11370`.
8. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL: `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.
9. Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2019. `arXiv:1906.02629`.
10. Lukáš Picek, Rafael Ruiz De Castañeda, Andrew M. Durso, Isabelle Bolon, and P. Mohanty Sharada. Overview of the snakeclef 2020: Automatic snake species identification challenge. In *CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece.*, 2020.
11. Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Weight standardization, 2019. `arXiv:1903.10520`.
12. 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, 2014. `arXiv:1409.0575`.

13. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. `arXiv:1312.6034`.
14. Yuxin Wu and Kaiming He. Group normalization, 2018. `arXiv:1803.08494`.
15. Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017. `arXiv:1710.09412`.