Spider Species Classification Using Vision Transformers and Convolutional Neural Networks

Arthur T. Magalhães
University of São Paulo - USP
Institute of Mathematics and Statistics
SP - São Paulo, Brazil
Email: amarthur@usp.br

Nina S. T. Hirata
University of São Paulo - USP
Institute of Mathematics and Statistics
SP - São Paulo, Brazil
Email: nina@ime.usp.br

***Abstract*—Spiders often seek shelter in the heat and safety of homes and although most of them are harmless, some can represent a real danger. Since differentiating spider species can be a challenge for individuals without prior knowledge, having a method to identify them could be useful in order to avoid potentially venomous ones. To address this question, this project aimed to analyze and compare the performance of convolutional neural networks (CNN) and vision transformers (ViT) regarding the quantitative and qualitative performance in the task of classifying different species of spiders from their images. We utilized publicly available images consisting of 25 Brazilian spider species and around 25,000 images. We selected the models based on their metrics and generalization performance in this classification task. The preliminary results indicated that ConvNeXt emerged as the most proficient among the examined Convolutional Neural Networks, achieving a macro accuracy of 88.5%. As for the Vision Transformers, MaxViT surpassed its counterparts, registering a macro accuracy of 90.1%, and outperformed the models in a direct comparison of their performance metrics. These results may contribute to the development of applications aimed at identifying spiders and providing information of interest about the species.**

I. Introduction

There are more than 48,000 known species of spider in the world, of which more than 4,500 are found in Brazil [1]. They are predators responsible for a fundamental role in controlling insect populations, which includes the biological control of pests in homes, gardens and plantations. However, because they can be found in places like closets, corners of walls, basements, garages and backyards, they can come into contact with humans and cause accidents. From 2017 to 2021, the Brazilian Ministry of Health's epidemiological bulletin [2] reported that there were more than 150,000 cases of accidents caused by spiders, ranking them third in the number of reported incidents involving venomous animals. For instance, the Butantan Institute[1] receives more than a hundred spiders annually, captured by people in and around their homes, who seek information about the type, the risks and what to do in the event of bites.

Given that the traditional method of taxonomic classification relies on specialized professionals, tools that can provide information about the spider species, habitat, potential harm, and other characteristics are an attractive alternative. This way, they could be useful for avoiding venomous species, possibly reducing the number of accidents as well as reducing the extermination of harmless spiders that are incorrectly thought to be dangerous. In this context, machine learning can be employed to automatically classify species, which could also be disseminated to the general public through a mobile app for example.

This idea is based on the success found in other applications in the field of Machine Learning [3], especially deep neural networks (Deep Learning) that have been proven to be efficient in processing complex, unstructured data such as images, video, text and audio [4]. In the case of image processing, we can think of neural networks as 'learning' during the training process to extract features and find suitable representations of the images being processed, which gives them the ability to generalize this knowledge to different images and distinct scenarios that they have never observed before. In this regard, many of the innovations in neural networks are architectural modifications that aim to improve this information extraction capacity. Currently, the two main architectures that stand out for their performance in the field of image processing are *Convolutional Neural Networks* (CNN) and *Vision Transformers* (ViT) [5], [6].

In our search, we have encountered only a limited number of studies addressing the challenge of spider recognition through these Deep Learning methods. In [7], authors consider 9 species and a dataset with 4478 samples, and report 90% accuracy on the validation set. In [8], authors also consider 9 species, however of species specific of Australia, having around 3,000 exemplars in each class. They report F1-score that ranges from 0.85 to 0.94. In [9], authors address the problem of spider sex recognition. Their dataset comprises 3,133 exemplars and report an accuracy of 92,38% on the validation set. However, it's worth noting that the datasets used in these studies are not publicly accessible. Therefore, we propose a different approach by utilizing publicly available images and restrict our attention to Brazilian spider species. Furthermore, we focus on two main aspects related to the networks: firstly, their ability to perform effectively in this task, and secondly, the potential performance disparities among them.

[1] https://butantan.gov.br/

II. Materials and Methods

A. Spiders Dataset

As a first step to begin the image classification process, we created a collection of spider photos from the iNaturalist website[2], which is an online platform supported by a global community dedicated to recording species observations. For the purposes of this work, we collected images of Brazilian spider species (order *Araneae*) with more than 100 photos and only those classified as *research grade*, a status obtained only when the community agrees with the identification, which ensures the consistency of our labels . Once the data collection stage was complete, we had 65 species with approximately 30,500 images. Naturally, there is a significant imbalance, influenced by a variety of factors including geographical and environmental conditions, human activity, visibility, among others. Because of this, we decided to utilize only a subset of our dataset, consisting of 25 species and 24,570 images, which make up roughly 80% of the total collected data (Fig. 1). Subsequently, we randomly partitioned 80% of the data into the training set and the remaining 20% into the validation set. It is worth mentioning that the application of the same training and validation sets across diverse networks ensures a consistent method for comparing the networks performance.

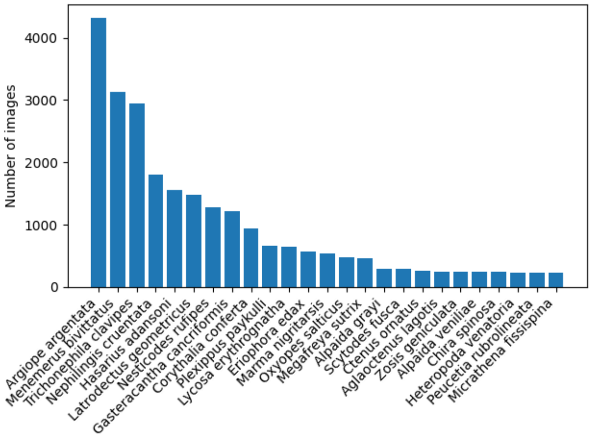


Fig. 1. Distribution of images in the selected subset of spider species.

Despite the substantial quantity of images, there exists significant variability in the dataset, encompassing aspects such as picture size, zoom, color, brightness and blur as well as natural factors like backgrounds and anatomical differences (Fig. 2). To address this variability, we execute a series of operations: resizing to ensure consistent input sizes; central cropping, which focuses on spiders, often found near the center and simultaneously removes background noise; and lastly, normalization to standardize pixel values in the image for consistent scale and distribution.

To further enhance the capabilities of our networks, we set up two methods to tackle the imbalance of species. Firstly, we apply a series of image augmentation transformations in the training set such as random cropping, flipping and rotation, which increases the diversity of the dataset by generating new samples (exemplified in Figure 3) and helps the model generalize better. Particularly, since spiders have a wide range of orientations given their specialized anatomy, this strategy can prove to be even more effective as the network learns to recognize them from various angles. Secondly, we used a weighted random sampler so that the data is sampled in a weighted way, which helps reducing bias in favor of majority classes and improves performance in the minority classes.

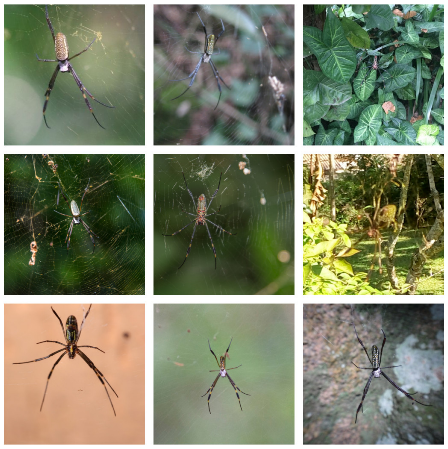
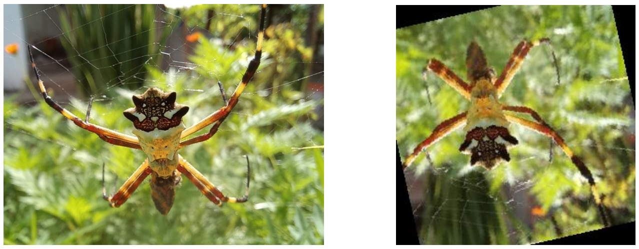


Fig. 2. Example of variations in images of *Trichonephila clavipes* spiders.



(a) Original image.

(b) A random transform.

Fig. 3. Example of a random transform for data augmentation.

B. The Neural Networks

To begin our analys, we selected architectures with modern design and proven effectiveness, evidenced by their performance on benchmark datasets such as the ImageNet dataset [10]. Specifically, we selected a set of state-of-the-art models that include ResNet, ResNeXt and ConvNeXt for the Convolutional Neural Networks and ViT, MaxViT and Swin for the Vision Transformers [6], [11]–[15].

As a starting point for training them, we employed an approach known as transfer learning, where a model that has already been pre-trained on another dataset is used as a basis for training a model on a related task. In the scope of this project, we used models pre-trained on millions of images from the aforementioned ImageNet dataset, which include a

[2] https://www.inaturalist.org/

variety of classes such as objects, vehicles, plants and even some spiders.

However, two modifications were required in order to make use of these models: the first one was to modify the network classifier and change the output to match the number of spider species classes that constitute the content of our data. The second one was to define which of the parameters would remain frozen and which ones would be trainable so that the model would be able to adapt to the specific characteristics of the new dataset. In the context of utilizing pre-trained models, features extracted from earlier layers tend to be more general, whereas those from later layers are more task-specific and the choice of whether or not to fine-tune the layers of the networks depends on the size of the target dataset and the layers' number of parameters [16]. Based on this, while using ImageNet as backbone, we unfreeze some of the parameters in the last layers while maintaining the remainder of the layers unaltered, given the constraint of our dataset's size relative to the network's scale and the limited common features shared with the original dataset. For all of our networks tested, we kept the number of trainable parameters about the same ($\approx$ 15.5M) to analyze generalization performance. In order to accomplish this, we utilized PyTorch framework which enabled us to use pre-trained models, adapt, train and validate our networks.

C. *Hyperparameters*

Before starting effectively the training phase, it was necessary to define some hyperparameters that, unlike the model parameters which are learned during the training process, they dictate the network's traits and learning behavior throughout the learning process, impacting its ability to generalize and thereby affecting its overall performance. Considering this, we employed random search to explore a predefined hyperparameter space, aiming to identify the optimal set of hyperparameters for each model within reasonable computation time [17].

- Loss Function: Cross Entropy Loss
- Learning Rate: $[10^{-4}, 10^{-1}]$ (log-uniformly)
- Batch Size: 32, 64, 128, 256
- Optimizers: Adam, AdamW
- Learning Rate Scheduler: Step Decay, Exponential Decay, Reduce On Plateau, Cosine Annealing

To ensure uniform evaluation across all models, we set the number of training epochs to a constant value of 30.

D. *Experiment Metrics*

To evaluate the performance of our networks, we utilized three key metrics: Micro Accuracy, Macro Accuracy, and Macro F1 Score. These metrics are able to provide a comprehensive understanding of model efficacy, while also functioning as reliable means of comparison. In light of this, we select the epoch that yields the highest macro accuracy on the validation set as the representative metric for model performance.

III. Results and Discussion

Following the implementation of the methodologies described, the results were tabulated and are summarized in Table I.

TABLE I
Accuracy and F1-score performance metrics across multiple models.

Model	Micro Acc (%)	Macro Acc (%)	Macro F1 (%)
Resnet50	87.30	87.77	85.71
ResNeXt	89.85	88.24	88.85
ConvNeXt	88.60	88.50	87.10
ViT	84.65	83.50	82.10
MaxViT	90.54	90.10	88.74
Swin	89.91	89.00	88.32

From this data, we can observe that Resnet50 has the lowest performance metrics among the tested CNNs, and the same can be observed on ViT for the Vision Transformers, although its metrics are significantly lower than any of the other models. By excluding these two lowest-performing models, we can obtain a more accurate representation of general performance through the mean analysis of the metrics, as illustrated in Table II. From it, two main observations can be drawn. Firstly, both architectures are able to perform effectively on this task. Secondly, the three metrics are closely aligned for all models, which suggests that the techniques we've applied to mitigate the class imbalance problem are effective and the models are able to perform consistently across the multiple classes.

TABLE II
Summary of best performance metrics for different model categories.

Metric	Category	Mean (%)	SD (%)
Micro Acc	CNNs	89.22	0.625
	ViTs	90.22	0.315
Macro Acc	CNNs	88.37	0.130
	ViTs	89.55	0.303
Macro F1	CNNs	87.98	0.875
	ViTs	88.53	0.210

Given that both architectures have good general capability, we can compare their individual models (Table I). In the CNN group, the ResNeXt model outperforms ConvNeXt in terms of validation performance. However, there is a substantial discrepancy of 10% between the training and validation metrics for ResNeXt, which suggests the presence of overfitting, even with regularization techniques. In contrast, the ConvNeXt model demonstrates a smaller generalization gap of 2%. The same phenomenon happens in the ViT group, where MaxViT outperforms Swin, and the generalization gap values are 4.5% and 4%, respectively. Given that the validation set might not be fully representative of the complete range of variance expected in real-world image data (especially for the classes with fewer images), choosing the models with a smaller generalization gaps could be more adequate as they are likely to have better generalization capabilities.

IV. Conclusion

The preliminary results indicated that both CNN and ViT are effective in the challenge of classifying spider species. Among the studied models, we take into consideration overall metric performance and generalization capabilities. In this regard, ConvNeXt (CNN) and MaxVit (ViT) have shown notable promise with macro accuracies of 88.5% and 90.1%. While the MaxVit outperforms the ConvNeXt in our tests, it would be premature to definitively state that one model is superior to the other. Experiments with a larger volume of data are needed to reach a more categorical conclusion. Further analysis is being conducted to broaden the dataset to encompass a greater number of spider species and employ other techniques such as ensemble methods to enhance network performance. This could yield more accurate outcomes, with implications for diverse real-world applications, including ecological studies, pest management, and other socially relevant fields.

Acknowledgment

The authors thank FAPESP (grants 2015/22308-2 and 2022/14286-2).

References

[1] Secretaria Municipal da Saúde, "Aranhas," 2020, https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/vigilancia_em_saude/controle_de_zoonoses/animais_sinantropicos/.

[2] Secretaria de Vigilância em Saúde, "Panorama dos acidentes causados por aranhas no brasil, de 2017 a 2021," 2022, https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2022/boletim-epidemiologico-vol-53-no31.

[3] Y. S. Abu-Mostafa, H.-T. Lin, and M. Magdon-Ismail, Learning From Data. AMLBook, 2012.

[4] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.

[5] Z. Li, W. Yang, S. Peng, and F. Liu, "A survey of convolutional neural networks: Analysis, applications, and prospects," 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR. OpenReview.net, 2021.

[7] Y. Jian, S. Peng, L. Zhenpeng, Z. Yu, Z. Chenggui, and Y. Zizhong, "Automatic classification of spider images in natural background," in 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), 2019, pp. 158–164.

[8] R. O. Sinnott, D. Yang, X. Ding, and Z. Ye, "Poisonous spider recognition through deep learning," in Proceedings of the Australasian Computer Science Week Multiconference, ser. ACSW '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3373017.3373031

[9] Q. Chen, Y. Ding, C. Liu, J. Liu, and T. He, "Research on spider sex recognition from images based on deep learning," IEEE Access, vol. 9, pp. 120 985–120 995, 2021.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[12] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," CoRR, vol. abs/1611.05431, 2016. [Online]. Available: http://arxiv.org/abs/1611.05431

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022.

[14] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," 2022.

[15] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," 2022.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" CoRR, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[17] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, 2012.