

Classificação de Espécies de Pássaros no Território Brasileiro com Base em Imagens e Metadados Geo-Temporais

Fabício Pereira Diniz¹, José Luis Seixas Junior¹, Rodrigo Clemente Thom de Souza^{2,3}

¹Ciência da Computação
Universidade Estadual do Paraná (UNESPAR)
Apucarana – Paraná

²Inteligência Artificial e Engenharia de Software
Universidade Federal do Paraná (UFPR)
Jandaia do Sul – Paraná

³Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Maringá (UEM)
Maringá – Paraná

Resumo. *Este trabalho propõe a classificação multimodal de aves em território brasileiro, utilizando dados de ciência cidadã. A pesquisa revelou que a distribuição dos registros é desigual entre os estados, refletindo a dependência da distribuição geográfica dos colaboradores. Observou-se que a acurácia dos modelos não demonstrou forte correlação com a quantidade de amostras. O desempenho variou significativamente com a modalidade de entrada. Os classificadores atingiram médias de acurácia de 0,8550 para dados tabulares, 0,3458 para imagens e 0,8606 para a combinação das modalidades. A concatenação dos embeddings demonstrou um ganho de acurácia, criando um espaço de características híbrido. Este espaço combinado a estabilidade dos metadados tabulares com a expressividade das imagens, resultando em menor sensibilidade à escolha do algoritmo de classificação. Em geral, o estudo estabelece uma baseline para a classificação multimodal de aves no Brasil e oferece subsídios para a aplicação de diferentes arquiteturas em problemas de Classificação Visual Fina.*

1. INTRODUÇÃO

A conversão de habitats naturais em mosaicos de uso antrópico, como agricultura e pastagens, afeta a riqueza de espécies, especialmente na Amazônia, onde a agricultura industrial é uma ameaça à biodiversidade [1, 2]. O conhecimento ornitológico no Brasil vem sendo impulsionado por iniciativas de ciência cidadã, com portais como o WikiAves utilizando fotografias como base taxonômica aviária [3]. A identificação automatizada de espécies é fundamental para a conservação ecológica, especialmente em cenários de perda de biodiversidade [4, 5]. Este trabalho propõe a classificação multimodal de aves no território nacional utilizando imagens e metadados geo-temporais, fundamentada na premissa de que a fusão de informações contextuais e visuais pode aprimorar a precisão [6]. Embora a ciência cidadã aumente a disponibilidade de amostras, os experimentos ainda dependem da distribuição geográfica dos colaboradores.

2. DESENVOLVIMENTO

O estudo adota uma abordagem quantitativa e experimental, baseada em arquiteturas Transformer. O Transformer, utilizando o mecanismo de Atenção, supera limitações de paralelização das Redes Neurais Recorrentes (RNNs) [7, 8]. Em visão computacional, o Vision Transformer (ViT) adapta essa lógica para modelar relações globais entre regiões de imagem [9, 10]. Para Classificação Visual Fina de Espécies (FGVC), o SwinFG é utilizado no branch de imagens devido ao seu mecanismo de janelas deslocadas, que restringe a Autoatenção a regiões menores enquanto captura relações de longo alcance [11]. Para os metadados tabulares, o TabTransformer foi empregado, o qual converte embeddings paramétricos em embeddings contextuais robustos, buscando aproximar o desempenho de modelos baseados em árvores, como o Random Forest [12, 13].

A metodologia utiliza uma estrutura dual-branch: os embeddings extraídos pelo SwinFG e TabTransformer são avaliados em três cenários: Tabular Isolado, Imagem Isolada e Concatenado (fusão). O conjunto de dados foi limitado às cinco espécies de maior ocorrência por estado para mitigar o desbalanceamento. Para a classificação final (downstream), foram testados diversos modelos, incluindo k-Nearest Neighbors (k-NN) [14], Random Forest [13], Support Vector Machine (SVM) [15], Logistic Regression (LogReg) [16] e XGBoost [17]. A métrica de avaliação utilizada em todos os cenários foi a Acurácia.

3. CONSIDERAÇÕES FINAIS

Os resultados demonstram que a quantidade de amostras não apresentou forte correlação com a acurácia. A maior variação de desempenho foi observada no domínio da Imagem. A acurácia média obtida foi de 0,8550 para a modalidade tabular, 0,3458 para a imagem e 0,8606 para a combinação das modalidades [13, 12]. A modalidade tabular, por si só, demonstrou alta estabilidade e consistência, enquanto a modalidade Imagem isolada apresentou o pior desempenho. A concatenação dos embeddings demonstrou ganho de acurácia em diversos cenários, criando um espaço de características híbrido que combina a estabilidade dos dados tabulares com a expressividade das imagens. Essa fusão multimodal mitigou a sensibilidade à escolha do algoritmo classificador downstream, apresentando maior robustez preditiva. A acurácia do classificador de imagem apresentou correlação negativa com a diversidade de espécies e positiva com a dominância de poucas espécies. O trabalho estabelece uma baseline para a classificação multimodal de aves no Brasil e fornece subsídios para a aplicação de diferentes arquiteturas em problemas de Classificação Visual Fina [11].

Referências

- [1] Lisieux Fuzessy, Sandrine Pavoine, Laura Cardador, Joan Maspons, and Daniel Sol. Loss of species and functions in a deforested megadiverse tropical forest. *Conservation Biology*, 38(4):e14250, 2024. doi: <https://doi.org/10.1111/cobi.14250>. URL <https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.14250>.
- [2] Beatriz Tavares Pinheiro, Sara Miranda Almeida, and Marcos Pérsio Dantas Santos. The impact of rare and common species on the functional diversity of forest birds in a palm-dominated landscape in the eastern amazon. *Acta Oecologica*,

126:104060, 2025. ISSN 1146-609X. doi: <https://doi.org/10.1016/j.actao.2025.104060>. URL <https://www.sciencedirect.com/science/article/pii/S1146609X25000049>.

- [3] José Fernando Pacheco, Luís Fábio Silveira, Alexandre Aleixo, Carlos Eduardo Agne, Glayson A. Bencke, Gustavo A. Bravo, Guilherme R. R. Brito, Mario Cohn-Haft, Giovanni Nachtigall Maurício, Luciano N. Naka, Fábio Olmos, Sérgio R. Posso, Alexander C. Lees, Luiz Fernando A. Figueiredo, Eduardo Carrano, Reinaldo C. Guedes, Evaldo Cesari, Ismael Franz, Fábio Schunck, and Vitor de Q. Piacentini. Annotated checklist of the birds of Brazil by the Brazilian ornithological records committee—second edition. *Ornithology Research*, 29(2):94–105, Jun 2021. ISSN 2662-673X. doi: 10.1007/s43388-021-00058-x. URL <https://doi.org/10.1007/s43388-021-00058-x>.
- [4] Javier Rodriguez-Juan, David Ortiz-Perez, Manuel Benavent-Lledo, David Mulero-Pérez, Pablo Ruiz-Ponce, Adrian Orihuela-Torres, Jose Garcia-Rodriguez, and Esther Sebastián-González. Visual wetlandbirds dataset: Bird species identification and behavior recognition in videos, 2025. URL <https://arxiv.org/abs/2501.08931>.
- [5] Javier Rodriguez-Juan, David Ortiz-Perez, Manuel Benavent-Lledo, David Mulero-Pérez, Pablo Ruiz-Ponce, Adrian Orihuela-Torres, Jose Garcia-Rodriguez, and Esther Sebastián-González. Visual wetlandbirds dataset: Bird species identification and behavior recognition in videos. *Scientific Data*, 12(1):1200, 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05516-5. URL <https://doi.org/10.1038/s41597-025-05516-5>.
- [6] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010. ISSN 1432-1882. doi: 10.1007/s00530-010-0182-0. URL <https://doi.org/10.1007/s00530-010-0182-0>.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [8] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0. URL <https://doi.org/10.1038/323533a0>.
- [9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, Hidenori Machino, Kazuma Kobayashi, Ken Asada, Masaaki Komatsu, Syuzo Kaneko, Masashi Sugiyama, and Ryuji Hamamoto. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1):84, September 2024. ISSN 1573-689X. doi: 10.1007/s10916-024-02105-8. URL <https://doi.org/10.1007/s10916-024-02105-8>.

1007/s10916-024-02105-8. Systematic review on ViTs vs CNNs in medical imaging.

- [11] Zhipeng Ma, Xiaoyu Wu, Anzhuo Chu, Lei Huang, and Zhiqiang Wei. Swinfg: A fine-grained recognition scheme based on swin transformer. *Expert Systems with Applications*, 244:123021, 2024. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2023.123021>. URL <https://www.sciencedirect.com/science/article/pii/S0957417423035236>.
- [12] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.
- [13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- [14] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.
- [15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL <https://doi.org/10.1007/BF00994018>.
- [16] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1958.tb00292.x. URL <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>.
- [17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL <https://doi.org/10.1145/2939672.2939785>.