Projeto de Pesquisa

Iniciação Científica

Classificação de espécies de aranha utilizando Transformers Visuais e Redes Convolucionais

Estudante

Arthur Teixeira Magalhães

Bacharelado em Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Orientadora

Nina S. T. Hirata

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

São Paulo, 8 de maio de 2023

Resumo

As aranhas procuram frequentemente abrigo no calor e segurança das residências e, embora a maioria delas seja inofensiva, algumas podem representar um perigo real. Assim, uma vez que diferenciar espécies de aranhas pode ser um desafio para pessoas sem conhecimento prévio, ter uma forma de identificá-las poderia ser útil para evitar aquelas que são venenosas. Para solucionar esta questão, neste projeto propomos estudar e comparar redes neurais convolucionais (CNN) e *vision transformers* (ViT) quanto ao desempenho na tarefa de reconhecimento de espécies de aranhas a partir de suas imagens. Estas técnicas efetuam a extração automática de características e aprendem representações de alto nível das imagens, sendo atualmente amplamente utilizadas, por exemplo, em detecção de objetos e classificação de imagens. Especificamente, objetivamos avaliar o potencial dessas redes para a tarefa de classificação de espécies de aranhas, incluindo comparações de desempenho quantitativas e qualitativas entre os dois tipos de redes. Para isso, utilizaremos o banco de dados de fotos de aracnídeos do INaturalist. Os resultados deste projeto poderão contribuir para o desenvolvimento de aplicativos voltados para a identificação de aranhas e fornecimento de informações de interesse sobre as espécies.

1 Introdução e Justificativa

No mundo, são conhecidas mais de 48.000 espécies de aranha, das quais mais de 4.500 se encontram no Brasil (Secretaria Municipal da Saúde, 2020). Elas são predadores responsáveis por um papel fundamental no controle de populações de insetos, o que inclui o controle biológico de pragas em casas, jardins e plantações. Entretanto, pelo fato de poderem ser encontradas em lugares como armários, cantos de paredes, porões, garagens e quintais, elas podem entrar em contato com humanos e causar acidentes. De 2017 a 2021, o boletim epidemiológico do Ministério da Saúde (Secretaria de Vigilância em Saúde, 2022) reportou que, no Brasil, houve mais de 150.000 casos de acidentes ocasionados por aranhas, classificando-as como o terceiro tipo de animal peçonhento quanto ao número de notificações. A título de exemplo, o Instituto Butantan[1] recebe um grande número de aranhas por ano, capturadas por pessoas em suas casas e arredores, que buscam informações sobre o tipo, os riscos e o que fazer em caso de picadas.

Dado que o método tradicional de classificação taxonômica depende de um profissional especializado, ferramentas que possam fornecer informações sobre a espécie da aranha, seu potencial de dano, habitat e outras características são uma alternativa atraente . Dessa forma, elas poderiam ser úteis para evitar espécies venenosas, possivelmente reduzindo a quantidade de acidentes, como também diminuir o extermínio de aranhas inofensivas, incorretamente julgadas peçonhentas. Nesse sentido, é possível utilizar o aprendizado de máquina para fazer tal classificação de espécies, que ainda poderia ser difundida para um público geral com um propósito educacional, por exemplo, através de um aplicativo de celular.

Essa ideia é baseada no sucesso encontrado em outras aplicações na área de *Machine Learning* (Abu-Mostafa et al., 2012), especialmente nas redes neurais profundas (*Deep Learning*). Elas têm se mostrado eficientes no processamento de dados complexos e não estruturados como imagens, vídeo, texto e áudio (Goodfellow et al., 2016). No caso de processamento de imagens, podemos pensar que durante o processo de treinamento as redes neurais "aprendem" a realizar a extração de *features* e encontram representações

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

adequadas das imagens sendo processadas, o que lhes conferem capacidade de generalizar esse "conhecimento" para novas imagens e diferentes cenários que nunca observaram antes. Nesse contexto, muitas das novidades em redes neurais são modificações na arquitetura que visam melhorar essa capacidade de extração de informação. Atualmente, duas arquiteturas se destacam pelo desempenho no campo de processamento de imagens: *Redes Convolucionais* e *Transformers* (Liu et al., 2021; Dosovitskiy et al., 2021; Vaswani et al., 2017).

As Redes Convolucionais (também denominadas *Convolutional Neural Networks ou CNNs*) têm sido amplamente utilizadas em diversas tarefas que envolvem imagens, tais como classificação, descrição, segmentação e detecção de objetos, cobrindo um grande espectro de aplicações, desde classificação de pragas de cultivo (Thenmozhi and Srinivasulu Reddy, 2019) até detecção de pneumonia (Varshni et al., 2019). Essas redes consistem em percorrer e aplicar filtros em cada região da imagem, horizontal e verticalmente, de modo a produzir uma nova representação da imagem original, que será alimentada em uma rede *Fully Connected* para gerar a classificação final (a ideia é que a rede aprenda os filtros e os pesos durante a etapa de treinamento). Comparada com uma rede neural padrão, essa arquitetura aperfeiçoa a capacidade de captura de características espaciais de uma imagem e, por isso, são um dos métodos mais comuns de classificação de imagens na área de *machine learning* atualmente.

Quanto à arquitetura de *Transformers* (Vaswani et al., 2017), ela foi fundada, a princípio, como base de redes neurais utilizadas no processamento de linguagem natural (NLP), como é o caso de BERT (Devlin et al., 2019) e GPT-3. Entretanto, também têm sido utilizada para processamento de imagens, como por exemplo o Vision Transformer (ViT) (Dosovitskiy et al., 2021) e o SWIN Transformer (Liu et al., 2021) que, por terem apresentado alto desempenho, têm recebido bastante destaque na comunidade de *machine learning* recentemente. Essa influência teve forte impacto, por exemplo, na medicina, como na classificação de ultra-som de mama (Gheflati and Rivaz, 2022) e, em um contexto mais recente, detecção de COVID-19 em radiografias do tórax (Chetoui and Akhloufi, 2022). Diferentemente das CNNs, por exemplo, o ViT baseia-se no princípio de *self-attention*. Explicando em termos gerais (Fig. 1), a imagem de entrada (A) é divida

em pedaços (B), que serão "tokenizados" e alimentados em um *transformer*. Então, esse *transformer* utiliza os mecanismos de atenção para gerar uma codificação (C) que passará por uma rede neural final, responsável por classificar a imagem (D). Dependendo da estrutura do banco de dados, o modelo de *transformers* pode superar em quase quatro vezes o desempenho das CNNs (Paul and Chen, 2021).

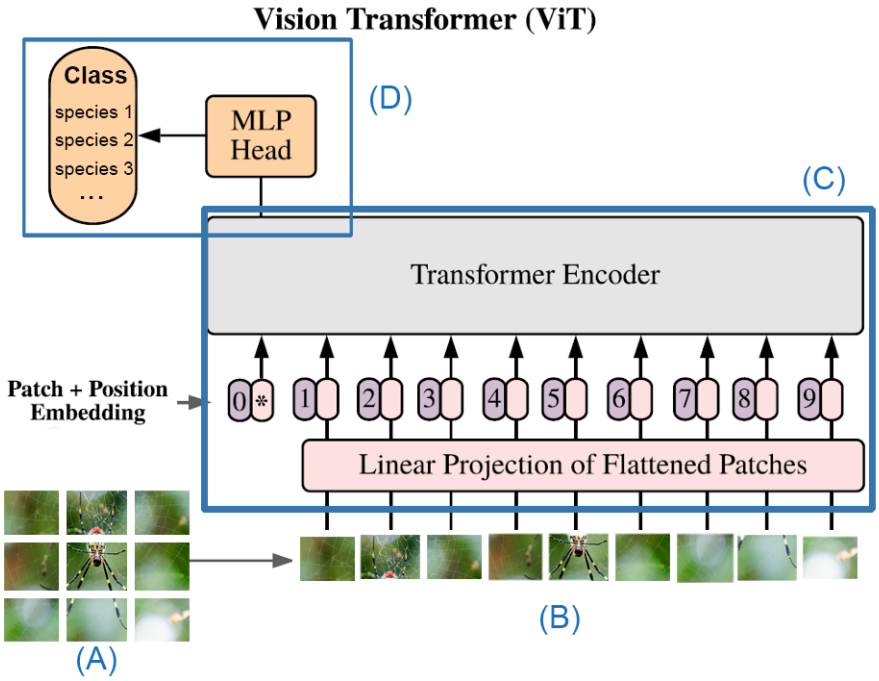

Figura 1: Representação do processo de classificação de um *vision tranformer* (ViT). Adaptado de Google Ai Blog.

Dessa forma, dada a importância da classificação de espécies de aranhas e o potencial de automatização dessa tarefa pelas redes neurais, duas perguntas surgem naturalmente: (1) essas redes são capazes de apresentar bom desempenho nessa tarefa? e (2) há diferença de desempenho entre esses dois tipos de redes e, em caso afirmativo, quais seriam elas? Um levantamento preliminar da literatura indica que há relativamente poucos trabalhos que abordam o reconhecimento ou classificação de espécies de aranhas e em geral esses trabalhos tratam um cenário bastante restrito. Por exemplo, Jian et al. (2019) testaram alguns modelos de CNNs usando cerca de 4500 imagens de 9 classes de aranhas, enquanto Chen et al. (2021) focam na discriminação do sexo de aranhas de uma certa espécie, usando imagens de alta resolução e um conjunto com cerca de 3000 imagens. Além disso, os conjuntos de imagens disponíveis publicamente também são limitados em termos de

quantidade de imagens e número de espécies.

2 Objetivos

Dado o contexto acima exposto, o principal objetivo deste projeto de pesquisa é responder as duas perguntas acima. Especificamente, queremos saber se as redes neurais e, em particular, as CNNs e os Vision Transformers são capazes de realizar a classificação de espécies de aranhas brasileiras de forma satisfatória, e como esses dois tipos de redes se comparam quanto ao desempenho nessa tarefa.

Objetivos específicos:

- Coletar fotos de espécies de aranhas brasileiras e construir um banco de dados de imagens, que serão utilizadas para treinar e validar as redes.
- Implementar as redes CNN e *transformer* por meio de bibliotecas de programação de Python.
- Treinar as duas redes, refinar seus hiper-parâmetros e coletar dados de seus comportamentos ao longo dessa etapa.
- Realizar uma análise comparativa detalhada das performances e das medidas estatísticas de desempenho entre as duas arquiteturas.

3 Material e métodos

3.1 Criação do Banco de Dados

Para iniciar o processo de criação do banco de dados, as imagens de espécies de aranhas brasileiras serão coletadas do site INaturalist (INat). Para garantir a integridade do banco de dados, utilizaremos somente imagens com licenças *Creative Commons* (CC) e categorizadas como *Research Grade*, isto é, quando uma observação verificável foi revista e a comunidade está de acordo com a identificação. As imagens serão analisadas e, caso seja necessário, poderão passar por uma etapa de pré-processamento e de normalização.

Em seguida, o conjunto de imagens será particionado nos subconjuntos de treinamento, de validação e, se houver uma quantidade suficientemente grande de imagens, de teste. Neste particionamento, um cuidado a ser tomado é garantir que a quantidade de classes tenha uma distribuição similar nas três partes para evitar viés nos resultados de classificação das redes.

3.2 Classificação das Imagens

A primeira etapa para iniciar a classificação de imagens será a construção do código das redes CNN e *transformer* em Python por meio da biblioteca *Pytorch*. Para tanto, utilizaremos redes pré-treinadas como ponto de partida, técnica conhecida como *transfer learning*: os parâmetros (pesos) da parte relacionada à extração de *features* serão incializados com valores obtidos em algum treinamento prévio que utiliza um conjunto de dados distinto (por exemplo, o bem conhecido ImageNet). Em seguida, essas redes deverão ser efetivamente treinadas sobre os conjuntos definidos anteriormente, em um processo conhecido como *fine-tuning*, no qual o ajuste fino dos hiper-parâmetros (*batch size*, *learning rate*, *epochs* etc) serão explorados, visando atingir as melhores performances. Resultados citados na literatura da área mostram que essa técnica reduz o tempo de treinamento além de melhorar a acurácia em geral sobre a nova tarefa-alvo.

Além disso, iremos monitorar eventuais ocorrências de *overfitting*, isto é, situação na qual o modelo se ajusta excessivamente bem no conjunto de treinamento ao mesmo tempo em que apresenta desempenho inferior em dados de teste. Esse monitoramento pode ser feito comparando-se as métricas de performance sobre o conjunto de treinamento e sobre o conjunto de validação ao longo do treinamento. Posto que, em geral, o *overfitting* ocorre devido à disparidade do número de parâmetros inerentes das redes comparado com o número de amostras de treinamento, pretendemos empregar técnicas de regularização como *data augmentation*, *dropout*, *L1* e *L2* para mitigar esse efeito. Em particular, *data augmentation* é uma ferramenta poderosa que consiste em aumentar o número de imagens artificialmente aplicando-se uma série de operações aleatórias como *position augmentation* (redimensionamento, corte, rotação, blur, zoom etc) e *color augmentation* (luminosidade, contraste, saturação etc) sobre as imagens originais. Esse aumento poderá ser relevante

especialmente para a classificação das espécies em que o número de imagens é reduzido.

Em razão da necessidade de recursos computacionais intensivos e de processamento por grandes períodos de tempo, utilizaremos as máquinas do Laboratório de Visão do IME-USP.

3.3 Avaliação dos Resultados

A princípio, a análise dos resultados será composta por quatro medidas estatísticas: *acurácia* (acc), *precisão* (prc), *recall* (rec) e *F1 score* (f1). Abaixo, estão exemplificados como esses parâmetros serão calculados, onde **VP** é verdadeiro positivo (imagens de uma classe classificadas como pertencentes a essa classe), **VN** é verdadeiro negativo (imagens que não pertenciam a uma classe e não foram classificadas como pertencentes a essa classe), **FP** é falso positivo (imagens classificadas como uma classe que não pertencem) e **FN** é falso negativo (imagens de uma classe detectadas como outra classe).

$$rec = \frac{VP}{VP + FN}$$

$$prc = \frac{VP}{VP + FP}$$

$$f1 = \frac{2 \times rec \times prc}{rec \ + prc}$$

$$acc = \frac{VP + VN}{VP + VN + FP + FN}$$

Embora os aspectos de interesse na comparação são claramente relacionados ao desempenho das redes na tarefa de reconhecimento de imagens, também analisaremos os aspectos computacionais como tempo de processamento e consumo de memória como forma de enriquecer a comparação.

Além de métricas quantitativas de desempenho, pretendemos fazer também análises qualitativas dos resultados. Em particular, pretendemos avaliar os mapas de ativação no caso de CNNs (Selvaraju et al., 2017) e os de atenção no caso de *transformers* (Chefer et al., 2021). Esses mapas indicam quais regiões da imagem mais afetam uma certa saída

da rede ou em quais pontos da imagem a rede está prestando mais atenção. Essa análise qualitativa poderá ser útil para explicar, em parte, eventuais diferenças de desempenho entre CNNs e *transformers*.

4 Plano de trabalho e cronograma de execução

A construção de classificadores baseados em CNN e em Transformers requer conhecimentos acerca de redes neurais e de conceitos e procedimentos relacionados ao treinamento e avaliação desse tipo de redes.

O estudante já cursou uma disciplina sobre Aprendizado de Máquina, na qual foi exposto aos fundamentos necessários para o desenvolvimento deste projeto, incluindo funcionamento e treinamento de redes neurais e métodos de seleção e/ou comparação de modelos. Além disso, desde setembro último vem participando do grupo de estudos, coordenado pela orientadora, que se reúne semanalmente para o estudo de *transformers*. Esses estudos já cobriram tópicos como redes recorrentes, redes recorrentes do tipo *encoder-decoder* para mapeamentos do tipo sequência para sequência, mecanismos de atenção e noções preliminares de *transformers*. No próximo mês esperamos concluir o estudo de *transformers* e *vision transformers*.

Conforme detalhado na seção anterior, o desenvolvimento do projeto envolverá três tarefas importantes, quais sejam: a construção de um banco de dados seguido por experimentação prática de como as arquiteturas se comportam e, por último, análise comparativa dos resultados. Neste sentido, este projeto contempla o processo completo de classificação de imagens por redes neurais. Por meio desse processo, espera-se que o estudante adquira conhecimentos relacionados à prática dos métodos usados em *Machine Learning*, e especificamente sobre o treinamento de CNNs e *transformers* para tarefas de classificação de imagens. Além disso, espera-se que as atividades de implementação e experimentação sejam acompanhadas de revisão e aprofundamento de conceitos e teorias que fundamentam esses métodos, ajudando o estudante a consolidar os seus conhecimentos sobre redes neurais e *machine learning*.

Adicionalmente, pretendemos também contemplar atividades que visam treinamento em investigação de literatura e em escrita científica. A investigação de literatura estará relacionada ao problema abordado (classificação de aranhas a partir de imagens usando técnicas de *deep learning*) e a escrita de textos científicos se dará por meio da elaboração de relatórios científicos e também de resumos e artigos a serem submetidos a eventos e, eventualmente, revistas da área.

Esse esquema de trabalho está representado no cronograma a seguir.

	Trimestre			
Atividade	1º	2º	3º	4º
Revisão e aprofundamento de fundamentos	**x**	**x**	**x**	**x**
Criação do banco de dados	**x**			
Estudo de trabalhos relacionados	**x**			
Treinamento das redes convolucional e *transformer*		**x**	**x**	
Análise dos resultados			**x**	**x**
Preparação do relatório final				**x**

5 Forma de análise dos resultados

O principal produto esperado deste projeto de pesquisa é um classificador para o reconhecimento de espécies de aranhas, a partir da imagem delas. Ou seja, a pesquisa a ser desenvolvida no projeto está relacionada à construção desse classificador. Estamos propondo o desenvolvimento de classificadores baseados em dois tipos de redes neurais, as CNNs e os *transformers*, os quais serão comparados e avaliados, para que seja escolhido um classificador final.

Os resultados da pesquisa serão avaliados em termos dos seguintes produtos ou atuações resultantes como decorrência direta da execução do projeto de pesquisa:

- Disponibilização pública do banco de dados de imagem e códigos desenvolvidos.
- Submissão ou publicação do trabalho em eventos ou revistas da área.
- Apresentação dos resultados em congressos e eventos científicos locais e regionais.
- Divulgação dos conhecimentos adquiridos através de atividades no ambiente univer-

sitário, como atividades em grupos de extensão e palestras.

Referências

Abu-Mostafa, Y. S., Lin, H.-T., and Magdon-Ismail, M. (2012). *Learning From Data*. AMLBook.

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, Los Alamitos, CA, USA. IEEE Computer Society.

Chen, Q., Ding, Y., Liu, C., Liu, J., and He, T. (2021). Research on spider sex recognition from images based on deep learning. *IEEE Access*, 9:120985–120995.

Chetoui, M. and Akhloufi, M. A. (2022). Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays. *J Clin Med*, 11(11).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

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

Gheflati, B. and Rivaz, H. (2022). Vision Transformers for Classification of Breast Ultrasound Images. *Annu Int Conf IEEE Eng Med Biol Soc*, 2022:480–483.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

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

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA. IEEE Computer Society.

Paul, S. and Chen, P.-Y. (2021). Vision transformers are robust learners.

Secretaria de Vigilância em Saúde (2022). Panorama dos acidentes causados por aranhas no brasil, de 2017 a 2021. `https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2022/boletim-epidemiologico-vol-53-no31`, Acessado em: 17 de Outubro de 2022.

Secretaria Municipal da Saúde (2020). Aranhas. https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/vigilancia_em_saude/controle_de_zoonoses/animais_sinantropicos/, Acessado em: 17 de Outubro de 2022.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Thenmozhi, K. and Srinivasulu Reddy, U. (2019). Crop pest classification based on deep convolutional neural network and transfer learning. *Computers and Electronics in Agriculture*, 164:104906.

Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., and Mittal, A. (2019). Pneumonia detection using cnn based feature extraction. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.