PROJETO FINAL XAI

Cesar A. P. Vial, Paulo S. A. Júnior

[1]CPGEI – Universidade Tecnológica Federal do Paraná (UTFPR)

Abstract. *ABSTRACT*

Resumo. *RESUMO*

1. Introdução

Com o aumento do uso de IA (Inteligência Artificial) ”caixa-preta” nos últimos anos nas mais diversas áreas, como as áreas de saúde, finanças, transporte e educação, a necessidade de explicar decisões tomadas por esses sistemas vem se tornando cada vez mais importante. Para tal, soluções de XAI (*eXplainable Artificial Intelligence*, Inteligência Artifical Explicável em inglês) vem sendo desenvolvidas de modo a suprir essa necessidade, garantindo aos usuários desses sistemas de IA mais confiabilidade e transparência [Dwivedi et al. 2023] e reduzindo os riscos em aplicações críticas, como saúde [Johannssen and Chukhrova 2025] e finanças [Arsenault et al. 2025].

O propósito da IA Explicativa é tornar o comportamento dos agentes de IA mais inteligíveis para humanos por meio da geração de explicações [Gunning et al. 2019] - um sistema XAI deve ser capaz de explicar suas capacidades, entendimentos, aquilo que foi feito, aquilo que está sendo feito, o que será feito depois e identificar quais informações estão sendo utilizadas por ele no momento [Bellotti and Edwards 2001]. Isso é feito pelas qualidades intrínsicas do modelo (por exemplo, árvores de decisão são intrinsicamente explicáveis) ou por meio de métodos *ante-* e *post-hoc*, que realizam análises sobre o agente de IA para gerar explicações quanto ao seu comportamento e resultados.

2. Trabalhos Recentes em XAI

Em [Kalasampath et al. 2025], os autores desenvolvem uma revisão literária sobre as aplicações de XAI no dia-a-dia em vários domínios. O trabalho explica alguns conceitos básicos: Transparência, que é definida como a medida de quanto usuários e *stakeholders* podem enxergar e compreender as operações internas do agente de IA, medido pela visibilidade em estrutura do modelo, movimento de dados e impacto de atributos individuais nos resultados finais, dando foco a como a transparência de um modelo pode aumentar a confiabilidade dele em campos críticos como saúde e finanças; e Interpretabilidade, que, de acordo com [Fan et al. 2021], mede o quanto uma pessoa consegue compreender e racionalizar modelos de IA, dividido em *simulabilidade* (compreensão sobre o modelo como um todo), *decomposibilidade* (compreensão sobre os componentes de um modelo) e *transparência algorítmica* (compreensão sobre o processo de treinamento e as dinâmicas de um modelo) - essas definições, porém, ainda são muito discutidas na literatura.

Na área da saúde, [Hulsen 2023] discute os conceitos e desafios do uso de IA, dando ênfase para os aspectos legais e éticos que acompanham esse tipo de aplicação (análise de imagens médicas, auxílio em diagnóstico e medicação, predição e automação de sistemas). O estudo aponta problemas como poucos dados causando enviesamento

em diagnósticos médicos, dificuldades no tratamento de privacidade e segurança em conformidade a regulamentações de proteção de dados como o GDPR (UE), HIPAA (EUA) e PIPL (China), e, especialmente no contexto de XAI, a confiança limitada que profissionais da saúde devem manter em sistemas de IA devido à sua natureza caixa-preta, problema cujas soluções devem manter em mente os problemas de enviesamento e privacidade em suas soluções.

Na área de finanças, [Arsenault et al. 2025] oferece uma visão geral dos usos de XAI no campo financeiro, apontando a crescente tendência de utilizar as técnicas de SHAP [Lundberg and Lee 2017] e LIME [Ribeiro et al. 2016], além de modelos com interpretabilidade intrínsica, para o desenvolvimento de sistemas de detecção de fraude, auxílio a decisões financeiras e preditores de *trends* do mercado.

Dentre as novas metodologias de XAI que estão sendo desenvolvidas, [He et al. 2025] explora o uso de interfaces de conversa como suplemento às técnicas de XAI, identificando uma melhora no entendimento de usuários sobre os sistemas de IA quando é possível estabelecer uma ”conversa” com o sistema. Porém, foi notado um aumento na dependência de usuários para com o sistema, proveniente do excesso de confiança que o usuário põe no sistema devido a explicações aparentemente bem-fundadas, mas ainda superficiais.

3. Proposta

Em [Maniparambil et al. 2024], são discutidas as diferenças entre *encoders* de texto e de imagens. O principal objetivo é verificar se *encoders* unimodais de linguagem e visão (imagens) possuem o mesmo entendimento do mundo, semanticamente, e verificar se essa similaridade no entendimento pode ser utilizada para relacionar ou explicar imagens sem treinamento adicional.

Para fazer isso, os autores utilizam um CKA (*Centered Kernel Alignment*) [Kornblith et al. 2019] para relacionar os resultados de ambos os *encoders*. No artigo original são utilizados vários *datasets* e modelos diferentes, para então verificar quais modelos possuem um resultado melhor após a aplicação do CKA, relacionando as diferentes descrições dos *embeddings* de texto, com as diferentes imagens dos *embeddings* de visão.

Neste artigo, a proposta principal é verificar a semelhança na explicabilidade/interpretabilidade nos dois modelos utilizados (de linguagem e visão). Serão aplicadas técnicas de XAI para ambos os modelos, com seus respectivos algorimos, para verificar se o foco nos objetos presentes nas imagens é similar ao foco em palavras nas descrições. O objetivo principal será na explicabilidade local, de acordo com as descrições e imagens relacionadas para um pequeno conjunto de imagens.

4. Experimentos e Resultados

4.1. Metodologia

Neste projeto, como o foco é na explicabilidade dos modelos, foi utilizado apenas um dataset, um modelo para texto e um modelo para imagens. O *dataset* utilizado foi o *COCO dataset* [Lin et al. 2015], que possui uma vasta base de imagens e de descrições genéricas para tais, possuindo um conjunto maior de figuras que descrições. O modelo utilizado

para a criação dos *embeddings* de texto foi o *RoBERTa* [Liu et al. 2019], e o modelo utilizado para a criação dos *embeddings* de imagens foi o *ResNet50* [He et al. 2015].

Para realizar os experimentos, foram agrupadas 5 imagens semelhantes e algumas das descrições consideradas compatíveis para tais imagens. O algoritmo *SHAP* foi utilizado para verificar quais as *features* (palavras) mais importantes nas descrições encontradas, e o algoritmo *GRAD-CAM* foi utilizado para a criação de mapas de calor indicando as áreas de foco nas imagens analisadas, de acordo com o resultado do modelo *ResNet50*. O esperado é que, para os melhores casos, as áreas focadas nas imagens estejam de acordo com as palavras mais importantes nas descrições.

O código do trabalho base foi utilizado para a criação dos *embeddings* e para calcular a relação das descrições com as imagens. Foi realizada uma implementação adicional para a aplicação dos algoritmos de XAI de acordo com esses resultados.

4.2. Resultados

Nesta seção, alguns resultados serão listados e discutidos. Todos os exemplos mostram um conjunto de 5 imagens, com as 2 descrições mais próximas delas. Abaixo de cada imagem está o seu mapa de calor, gerado pelo algoritmo *GRAD-CAM*, mostrando as regiões de foco de cada imagem. Junto com essa figura, terão gráficos do resultado da execução do algoritmo *SHAP* para cada uma das frases.

A figura 1 mostra um conjunto de imagens, nas quais 3 possuem mulheres jogando tênis, e as outras 2 possuem homens jogando tênis. Verificando os mapas de calor, o foco em todas as imagens parece ser o mesmo, no jogador e na raquete de tênis. Analisando os valores de *SHAP* nas figuras 2 e 3, podemos verificar uma grande influência das palavras *tennis*, *player*, *woman*, *racquet*, *court* e *ball*, todas aparecendo nos pontos focais das imagens analisadas.

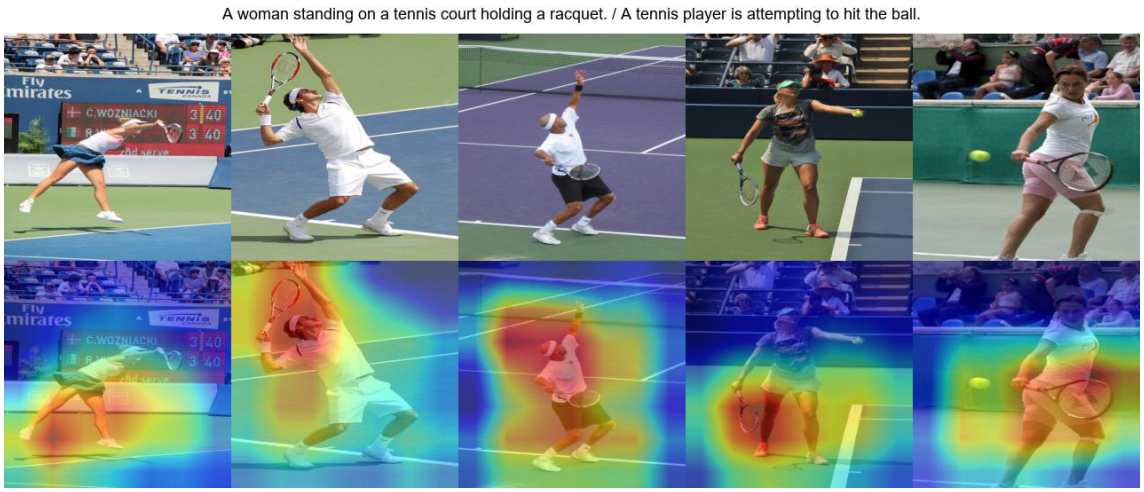

Figure 1. Conjunto de imagens com as mesmas descrições: *A woman standing on a tennis court holding a racquet. / A tennis player is attempting to hit the ball.*; Junto com a execução do algoritmo GRAD-CAM.

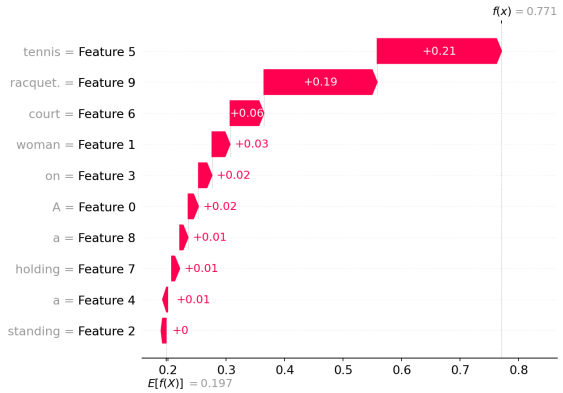

Figure 2. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A woman standing on a tennis court holding a racquet.*

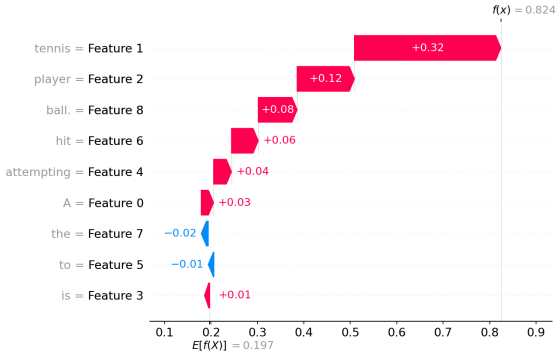

Figure 3. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A tennis player is attempting to hit the ball.*

A figura 4 mostra um conjunto de imagens com diferentes animais próximos a algum grande corpo de água, seja na beira de um rio, ou em cima de uma ponte. Em todas as imagens o foco é no grupo de animais. As figuras 5 e 6 mostram o foco nas palavras *river*, *lake*, *goats*, *herds*, *animals* e *land*. O foco das imagens é primariamente nos grupos de animais, sem muito foco nos corpos de água.

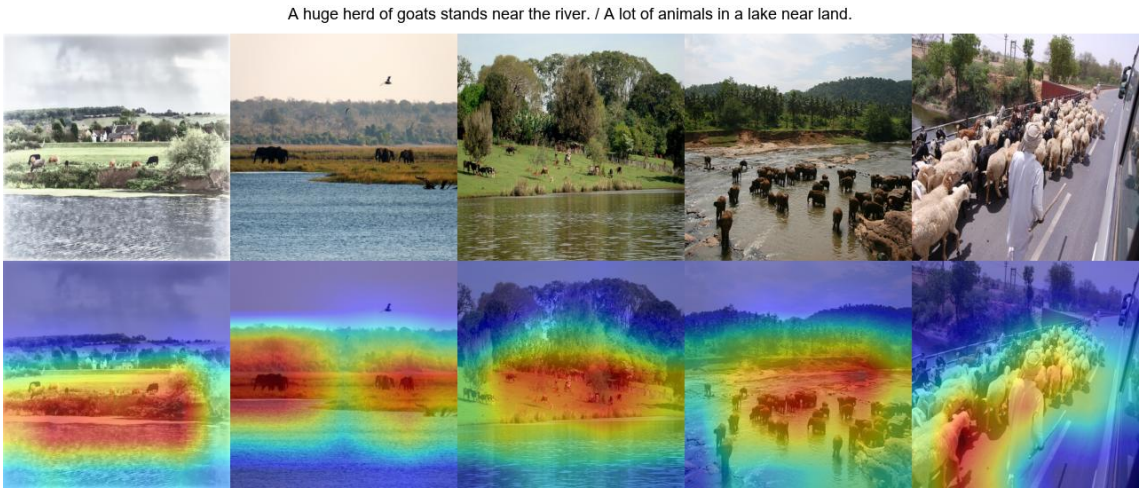

Figure 4. Conjunto de imagens com as mesmas descrições: *A huge herd of goats stands near the river. / A lot of animals in a lake near land.*; Junto com a execução do algoritmo GRAD-CAM.

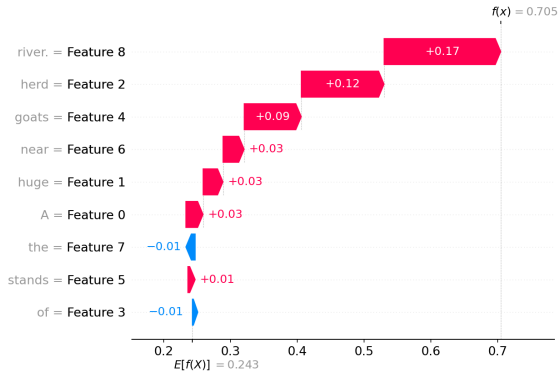

Figure 5. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A huge herd of goats stands near the river.*

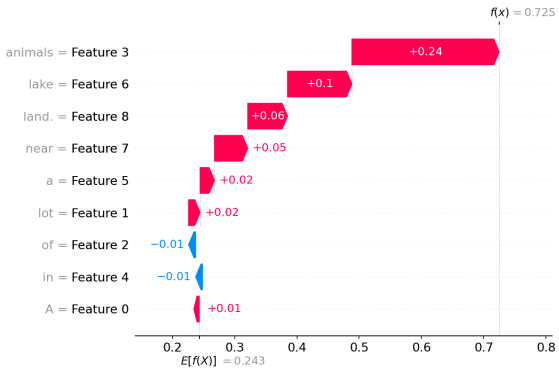

Figure 6. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A lot of animals in a lake near land.*

A figura 7 mostra múltiplas imagens com gatos e cachorros, dentro de casas. Na primeira figura, o foco ficou completamente fora do gato da foto, já as outras imagens todas possuem foco nos animais da foto. As figuras 8 e 9 mostram em ambas as descrições um foco absurdo na palavra *cat*, com um foco menor em *couch*, *laying* e *blankets*, em relação a posição e o local do gato. Mas como é possível ver, uma das imagens possui um cachorro, e em uma posição completamente diferente da descrita.

Figure 7. Conjunto de imagens com as mesmas descrições: *A cat that is laying down on a couch. / There is a cat that is laying on top of blankets.*; Junto com a execução do algoritmo GRAD-CAM.

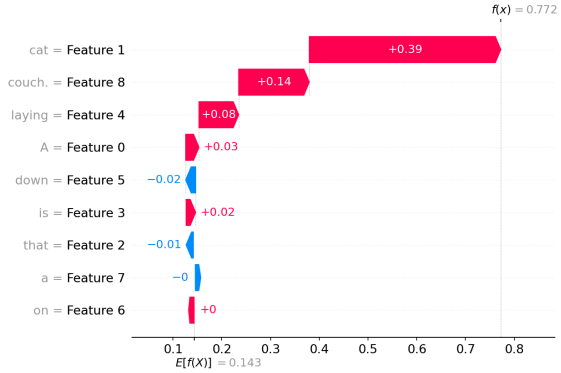

Figure 8. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A cat that is laying down on a couch.*

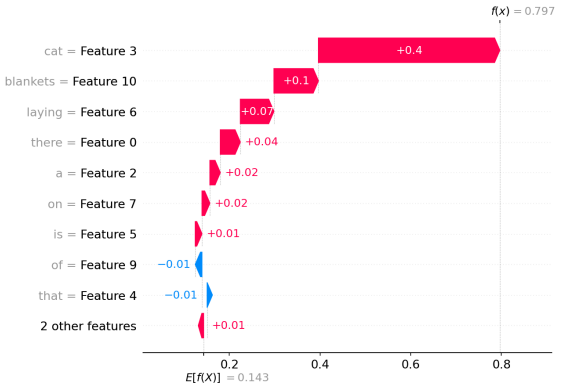

Figure 9. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *There is a cat that is laying on top of blankets.*

A figura 10 mostra um conjunto de imagens de rua, sendo três do lado externo no meio de ruas, uma em um mêtro, e uma no que parece ser um mercado. As descrições nas figuras 11 e 12 mostram uma grande quantidade de substantivos, e um foco parecido em parte delas. Um resultado interessante foi o peso negativo na palavra *horse*, possivelmente por não se encaixar na descrição de uma rua cheia de placas e carros. E isso se encaixa também com as imagens relacionadas, que mostram diferentes cenários urbanos, onde nenhuma dela possui um cavalo.

Figure 10. Conjunto de imagens com as mesmas descrições: *The image of a street with parked cars on the side walk and there are advertsment signs. / A street filled with different signs, cars, cyclist, and horse.*; Junto com a execução do algoritmo GRAD-CAM.

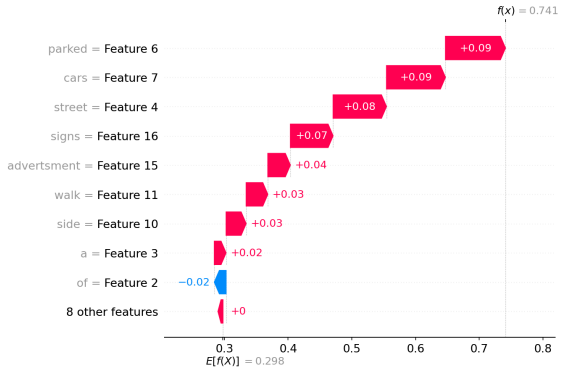

Figure 11. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *The image of a street with parked cars on the side walk and there are advertsment signs.*

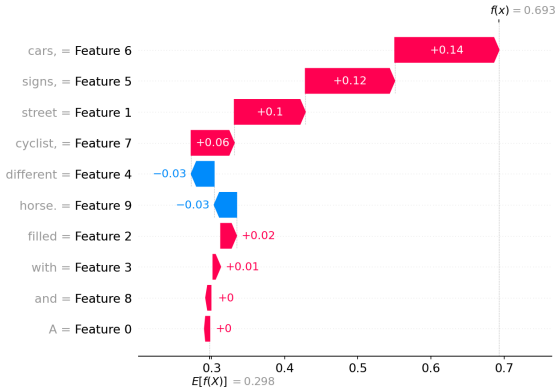

Figure 12. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A street filled with different signs, cars, cyclist, and horse.*

A figura 13 mostra um conjunto de imagens relacionadas a comidas, nas quais 3 delas mostram cachorros-quentes. Em todas as imagens, a comida, ou parte dela, parece ser o ponto de maior importância. De acordo com as descrições e a importância das palavras *vegetables*, *apples*, *kitchen* e *fruits*, como visto nas figuras 14 e 15, é possível afirmar que o modelo de agregação de *embeddings* não obteve um bom resultado para essas figuras e descrições. A única semelhança está no foco em comidas, tanto nas descrições quanto nas imagens.

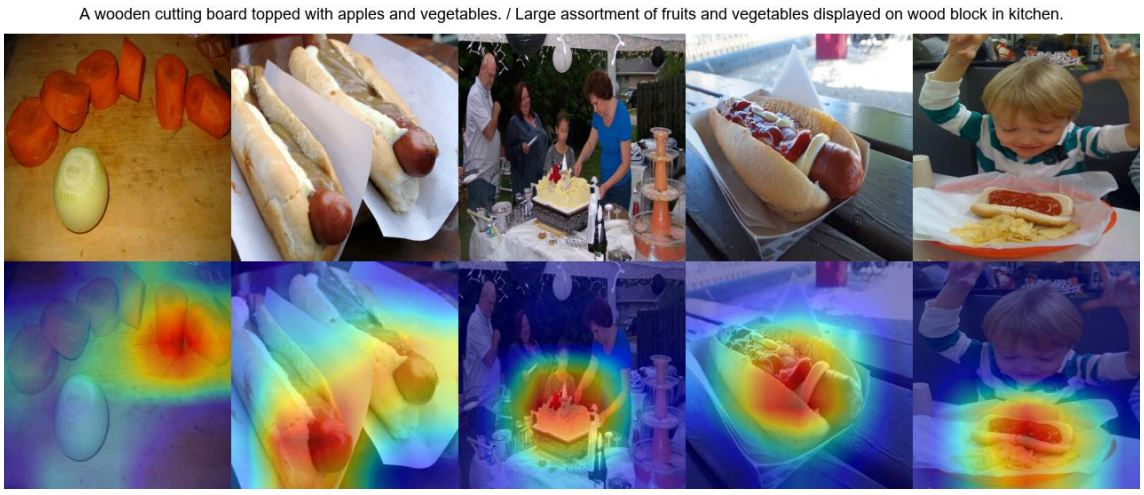

Figure 13. Conjunto de imagens com as mesmas descrições: *A wooden cutting board topped with apples and vegetables. / Large assortment of fruits and vegetables displayed on wood block in kitchen.*; Junto com a execução do algoritmo GRAD-CAM.

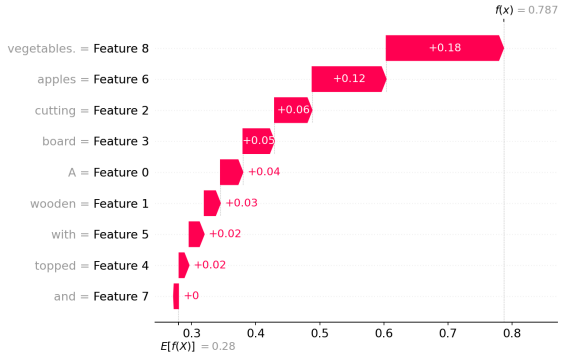

Figure 14. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A wooden cutting board topped with apples and vegetables.*

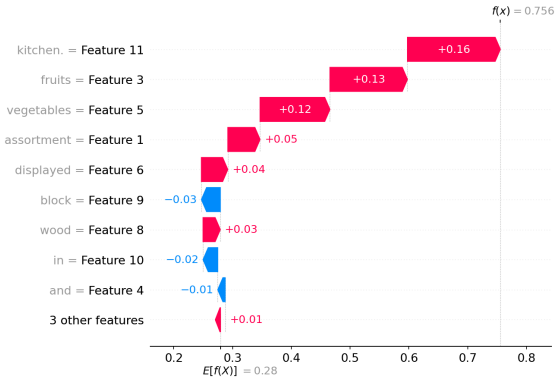

Figure 15. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *Large assortment of fruits and vegetables displayed on wood block in kitchen.*

5. Conclusão

Nos experimentos, foi possível verificar alguns padrões na execução do *GRAD-CAM*, com focos de acordo com a descrição relacionada. Já na aplicação do *SHAP* para detectar as palavras mais importantes em cada frase classificada, é possível inferir que substantivos e adjetivos normalmente possuem uma importância maior nas frases. Este comportamento corrobora com o que queremos encontrar na imagem, já que normalmente nas frases temos os objetos de maior importância (como um animal, pessoa, objeto ou lugar), muitas vezes com suas características. No geral, as explicações para frases e imagens foram similares, podendo se perder nos casos em que o modelo de agregação de *embeddings* encontrou uma boa relação entre o conjunto de imagens e as descrições.

6. Trabalhos Futuros

A pesquisa em XAI atualmente se direciona especialmente à criação de *benchmarks* e métricas para a avaliação das diferentes técnicas de XAI, possibilitando a escolha informada de determinada técnica ou outra para determinada aplicação. Além disso, o *overhead* computacional de certas técnicas mais populares (como o LIME) tem se mostrado um dos maiores problemas da área, e a redução de seu custo computacional é bastante interessante para os usuários de XAI. Para facilitar e popularizar o uso dessas técnicas, é interessante a proposição de *frameworks* de XAI que possam ser utilizados por usuários leigos junto às aplicações mais populares, como em *chatbots* e sistemas de predição. Por outro lado, ainda há muito trabalho a ser feito no quesito de integração de *guidelines* éticos às técnicas de XAI, além do mapeamento de áreas críticas na indústria que devem se beneficiar de seu uso e do alinhamento de itens das leis gerais de proteção de dados, como o LGPD e o GPDR, que devem ser mantidos em mente durante o uso de XAIs.

Para evoluir os estudos buscando explicar a relação entre textos e imagens, podem ser realizados testes utilizando difrerentes algoritmos na criação dos *embeddings*, visando melhores relações entre a descrição e a figura, assim como diferentes algoritmos de XAI para ambos os *embeddings*, analisando como eles classificam as palavras de

maior importância com as regiões de maior importância em cada imagem. Todos estes testes também poderiam ser relacionados utilizando um modelo que transforma imagens em texto, e vice-versa (modelos generativos ou de busca), para analizar quais as *features* mais importantes em figuras e frases para que o modelo possa gerar/encontrar a frase ou figura de acordo. Por fim, é possível executar o *Grad-CAM* utilizando as descrições relacionadas, no lugar do próprio modelo utilizado na criação dos *embeddings*, para verificar se o modelo poderia encontrar a descrição encontrada para tal imagem;

7. Acknowledgements

A ferramenta do *Chat-GPT* foi utilizada como suporte na implementação dos algoritmos citados.

References

[Arsenault et al. 2025] Arsenault, P.-D., Wang, S., and Patenaude, J.-M. (2025). A survey of explainable artificial intelligence (xai) in financial time series forecasting. *ACM Computing Surveys*, 57(10):1–37.

[Bellotti and Edwards 2001] Bellotti, V. and Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction*, 16(2-4):193–212.

[Dwivedi et al. 2023] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33.

[Fan et al. 2021] Fan, F.-L., Xiong, J., Li, M., and Wang, G. (2021). On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760.

[Gunning et al. 2019] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

[He et al. 2025] He, G., Aishwarya, N., and Gadiraju, U. (2025). Is conversational xai all you need? human-ai decision making with a conversational xai assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 907–924.

[He et al. 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[Hulsen 2023] Hulsen, T. (2023). Explainable artificial intelligence (xai): Concepts and challenges in healthcare. *AI*, 4(3):652–666.

[Johannssen and Chukhrova 2025] Johannssen, A. and Chukhrova, N. (2025). The crucial role of explainable artificial intelligence (xai) in improving health care management. *Health Care Management Science*, pages 1–6.

[Kalasampath et al. 2025] Kalasampath, K., Spoorthi, K., Sajeev, S., Kuppa, S. S., Ajay, K., and Angulakshmi, M. (2025). A literature review on applications of explainable artificial intelligence (xai). *IEEE Access*.

[Kornblith et al. 2019] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited.

[Lin et al. 2015] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

[Liu et al. 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

[Lundberg and Lee 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Curran Associates, Inc.

[Maniparambil et al. 2024] Maniparambil, M., Akshulakov, R., Djilali, Y. A. D., El Amine Seddik, M., Narayan, S., Mangalam, K., and O'Connor, N. E. (2024). Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14334–14343.

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.