# Utilização de XAI para explicação de encoders de imagem e texto

**Cesar A. P. Vial, Paulo S. A. Júnior**

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

## 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.

Outras áreas citadas por [Kalasampath et al. 2025] são a área de Educação, para geração de explicações para auxílio ao ensino e análise de dados vocacionais, a área de Cibersegurança, para detecção de *malware* e inspeção de tráfego de rede, a área de Direito, para inerpretação de casos, análise de contratos e auxílio a decisões judiciárias, e a área de Agricultura, para recomendação de plantios e análise de fatores de qualidade.

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*, alinhando um conjunto de imagens a um conjunto de descrições. 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

Para entender a metodologia do projeto, é preciso entender melhor o projeto base [Maniparambil et al. 2024] utilizado. Primeiramente, foi utilizado 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, de maneira que uma descrição pode servir para múltiplas imagens. Ressaltando que no artigo original várias bases de dados e modelos são utilizados, mas para este trabalho será utilizado apenas este *dataset*, um modelo para imagens e outro para texto.

Para a criação dos *embeddings* de imagem foi utilizado o modelo *ResNet50* [He et al. 2015]. Este modelo é um classificador, então dada uma imagem, ele retorna probabilidades da imagem dar foco para um certo objeto, utilizando as classes do próprio modelo do *ResNet50* do *torchvision*. As probabilidades retornadas pelo modelo podem ser utilizados como uma base para conferir a eficiência dessas classificações. Para a criação dos *embeddings* de texto, foi utilizado o modelo *RoBERTa* [Liu et al. 2019]. Este modelo apenas cria os embeddings, de acordo com os *tokens* (palavras) encontrados em cada frase. Utilizando os algoritmos de XAI durante a metodologia, é possível verificar quais são os tokens mais importantes, e dessa forma relacionar as diferentes frases próximas nos *embeddings*.

Com os dois *embeddings criados*, de maneira independente, o CKA é utilizado para aproximar ambos os *embeddings*. Este processo envolve uma matemática complexa, que foge do escopo deste trabalho. O artigo original entrega apenas resultados globais de cada modelo e *dataset* utilizado, logo não foram utilizadas métricas quantitativas para este passo do algoritmo. O que foi utilizado foi a presença dos *tokens* da criação dos *embeddings* de texto na classificação do modelo de imagens.

Para realizar os experimentos utilizando os algoritmos de XAI, foram agrupadas 5 imagens semelhantes e algumas das descrições consideradas compatíveis para tais imagens. O algoritmo *SHAP* foi utilizado para verificar quais os *tokens* 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 a classificação com maior probabilidade do modelo *ResNet50*. O esperado é que, para os melhores casos, as áreas focadas nas imagens estejam de acordo com a classificação, tal qual deve estar nas frases selecionadas. Este resultado implicaria no sucesso do algoritmo do projeto original, e a análise de XAI pode ser realizada para verificar se a explicabilidade dos modelos também é condizente entre ambos.

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. O código final está presente no repositório `https://github.com/cesarvial/0-shot-llm-vision`.

### 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 zebras em condições diversas. As descrições possuem relação com a previsão do ResNet50, que classificou as zebras na imagem com certeza. Esse é um exemplo do alinhamento correto entre os *embeddings* de *caption* e de imagem, e os métodos de XAI também refletem a corretude desse alinhamento - tanto nas figuras 2 e 3 como na figura 1 as *features* destacadas são principalmente as *zebras*, seguido de *grass* e *field*.


**Figure 1. Conjunto de imagens com as mesmas descrições: *A zebra looking up as another grazes in a field. / Two zebra's standing in a grassy field and one is eating grass.*; Junto com a execução do algoritmo GRAD-CAM e a predição com maior probabilidade do ResNet50: *zebra*.**

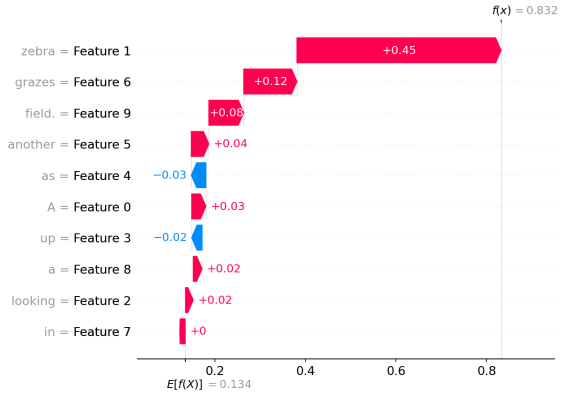


**Figure 2. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A zebra looking up as another grazes in a field.***

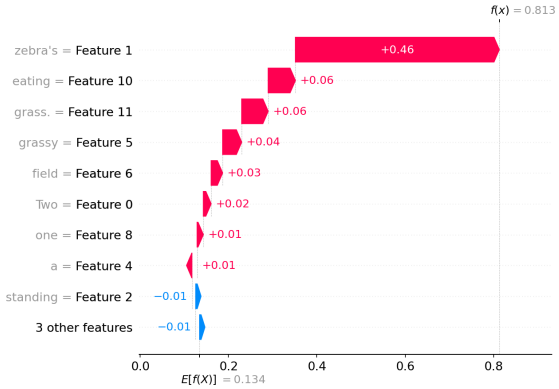


**Figure 3. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *Two zebra's standing in a grassy field and one is eating grass.***

Outro bom exemplo é dado na figura 4, que mostra jogadores de Tênis no processo de atingir a bola. Neste exemplo, o modelo ResNet50 não dá certeza absoluta à raquete, considerando também o jogador e o campo, enquanto o GradCAM da figura 4 aponta principalmente para os jogadores e para as raquetes, além do campo, enquanto nas figuras 5 e 6 a ênfase foi dada às *features tennis*, *player*, *court* e *ball*.

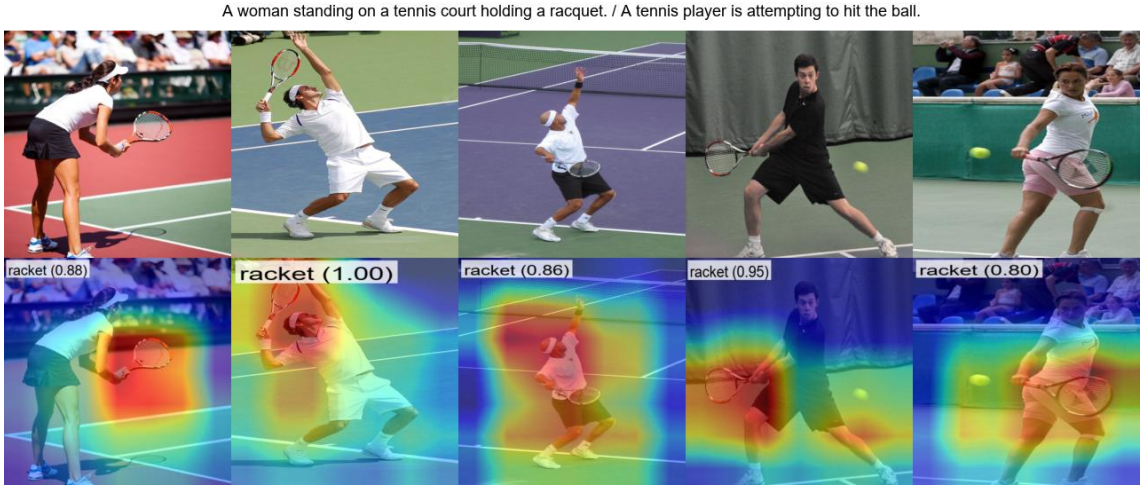


**Figure 4. Conjunto de imagens com as mesmas descrições: *A woman standing on a tennis court holding a racquet. / A tennis player is attempting to hit the ball.*; Junto com a execução do algoritmo GRAD-CAM e a predição com maior probabilidade do ResNet50: *racket*.**

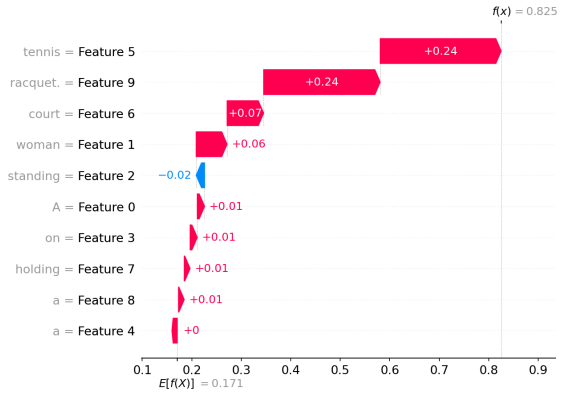


**Figure 5. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A woman standing on a tennis court holding a racquet.***

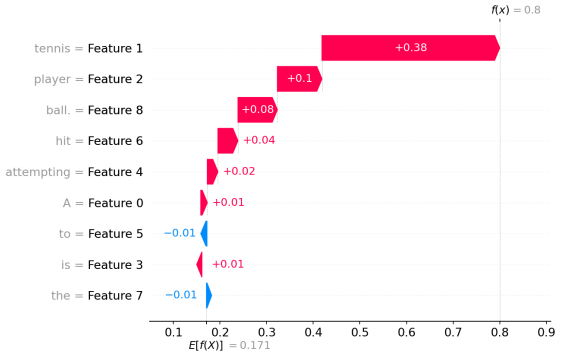


**Figure 6. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A tennis player is attempting to hit the ball.***

A figura 7 mostra 4 imagens com um sanduíche ou *hotdogs* em pratos, com uma imagem de uma menina comendo uma pizza. Para o *ResNet50* com o GRAD-CAM, o foco em todas as imagens é o alimento, com uma classificação de *hotdog* para todos eles. As descrições referenciam apenas sanduíches que, como visto pelas figuras 8 e 9, é a

palavra com maior foco, junto de *plate* e *salad*. Nenhuma imagem possui a palavra *sandwich* como uma classe, porém todas possuem múltiplas classes de tipos de sanduíches, como *cheeseburger, burrito, bagel* e até mesmo *pizza* (mostrado na última imagem). Isso mostra um boa performance da relação entre os *embeddings*, que mesmo que os *tokens* não sejam exatos, eles estão relacionados, tal relação é encontrada pelo algoritmo e reiterada pelas explicações dos modelos de XAI.


**Figure 7. Conjunto de imagens com as mesmas descrições: *A sandwich that is on a plate on a table. / a sandwich cut in half and a salad.*; Junto com a execução do algoritmo GRAD-CAM e a predição com maior probabilidade do ResNet50: *hotdog*.**

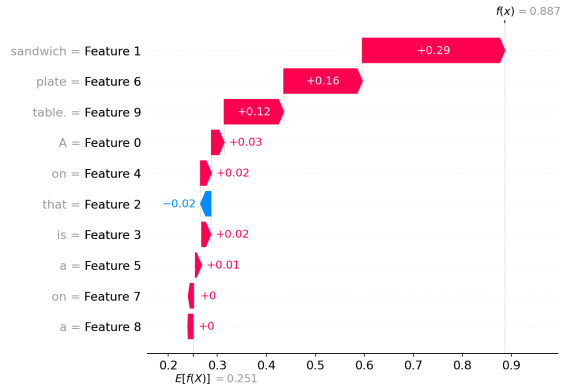


**Figure 8. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A sandwich that is on a plate on a table.***

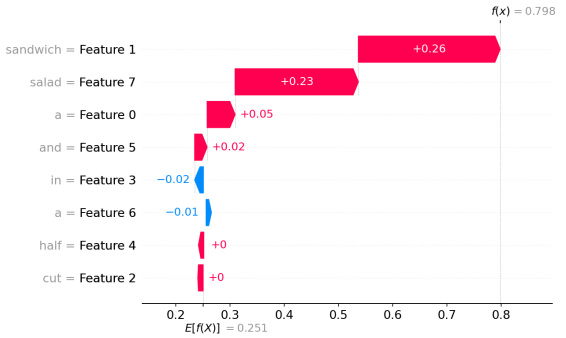


**Figure 9. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *a sandwich cut in half and a salad.***

A figura 10 mostra quatro imagens de bolo com uma imagem de um porta-lápis. Esse é um exemplo interessante por conta das classificações secundárias das imagens. As descrições possuem relação com as classificações gerais das imagens (com os *tokens*

*cupcake*, *plate* e *spoon*), sendo que as imagens de 1 a 4 as classificações *cake, plate e spoon* aparecem não com a maior probalidade, sendo essa dada para outros objetos na imagem. A imagem 5 está presente neste conjunto por conta de sua confusão do porta-lápis por *wall-clock* ou *analog-clock*, que são duas classificações alternativas para a imagem 1, que mostra um bolo no formato de relógio. Neste caso, o erro está no modelo de classificação de imagens, e não no algoritmo de relação de *embeddings*. As figuras 11 e 12 mostram a maior importância nas palavras *cupcake* e *plate*, que não são classificações primárias, mas são o foco das imagens, para praticamente todos os casos.


**Figure 10. Conjunto de imagens com as mesmas descrições: *A frosted cupcake and a silver spoon siting on a white plate. / A black-and-white image of a cupcake and spoon on a small plate*; Junto com a execução do algoritmo GRAD-CAM e a predição com maior probabilidade do ResNet50: *puck, candle, tray, matchstick, wall clock*.**

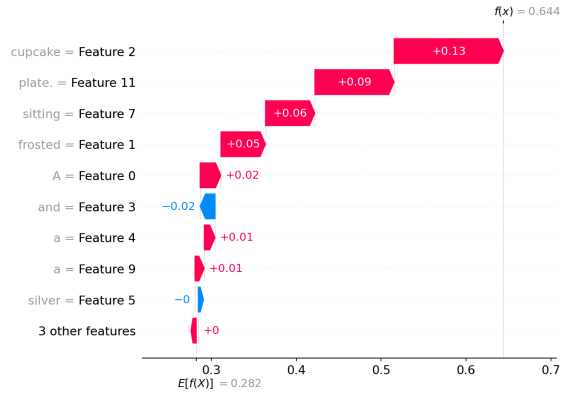


**Figure 11. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A frosted cupcake and a silver spoon siting on a white plate.***

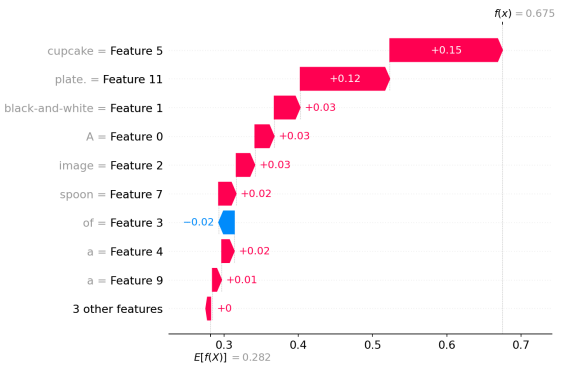


**Figure 12. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A black-and-white image of a cupcake and spoon on a small plate.***

A figura 13 mostra um conjunto de imagens de salas de jantar e cozinhas. O modelo *ResNet50* configura a maior parte dessas imagens como restaurantes, com o foco do GRAD-CAM nas mesas para tais casos. Uma das imagens possui a classificação de *patio* com probabilidade baixa, e a única relação está na quinta maior probabilidade para tal imagem, sendo *dining table*, com uma probabilidade de 0.0916. A última imagem mostra o foco nos balcões de uma cozinha, com a classificação *dishwasher*. As descrições fogem bastante das classificações primárias de cada imagem, aparecendo apenas na classe *dining table*, que é uma classe secundária para as imagens 1 até a 4. Por conta das baixas probabilidades nas classes principais e nenhuma delas fazendo parte dos tokens das descrições relacionadas, esse pode ser um exemplo no qual o algoritmo falhou, apesar de a explicabilidade fazer sentido para algumas imagens, pela presença de cozinhas, mesas e cadeiras que, como visto nas figuras 14 e 15, as palavras *kitchen, room, table, chair* possuem uma grande importância para as descrições.


**Figure 13. Conjunto de imagens com as mesmas descrições: *A room with a couch a tv near a kitchen / A chair and a table in a small room.*; Junto com a execução do algoritmo GRAD-CAM e a predição com maior probabilidade do ResNet50: *restaurant*, *patio* e *dishwasher*.**

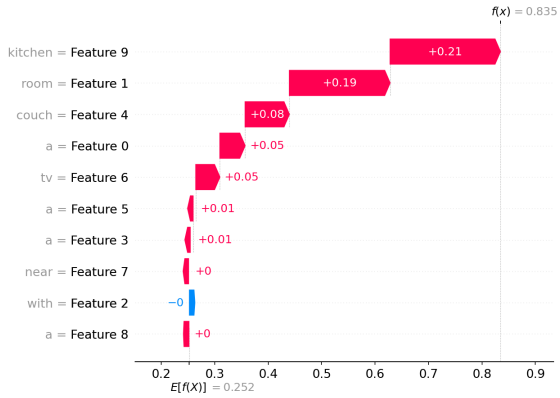


**Figure 14. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A room with a couch a tv near a kitchen.***

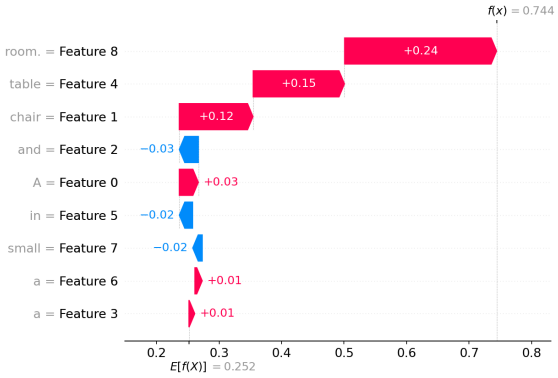


**Figure 15. Gráfico em cascata do resultado do algoritmo de SHAP para a descrição: *A chair and a table in a small room.***

## 5. Conclusão

Nos experimentos, foram analisados casos onde ambos os modelos e o algositmo de relação de *embeddings* teve um bom desempenho, assim como casos nos quais algum dos passos falhou. Neles, 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.