



Projeto Final XAI

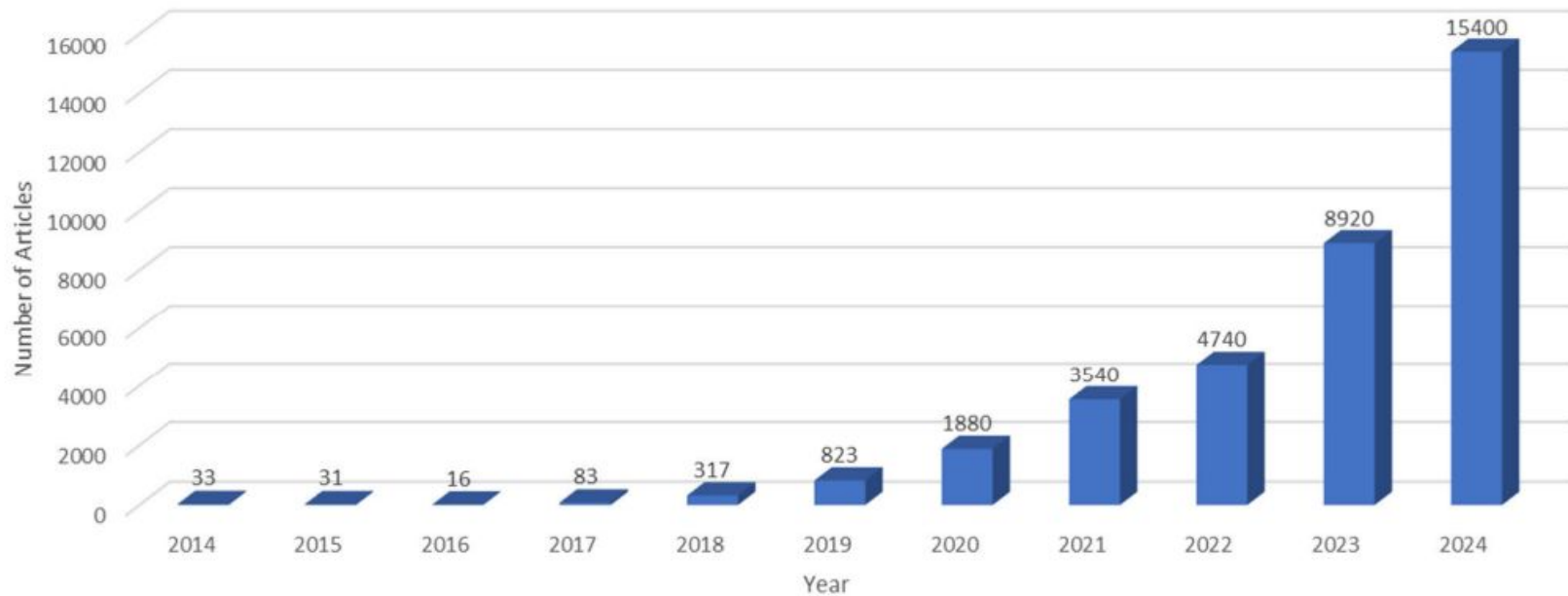
Cesar Augusto Pereira Vial
Paulo Sérgio Ávila Júnior



Introdução

O advento do XAI nos últimos anos tem se dado pela:

- Popularização do uso de métodos “Caixa-Preta”, como as LLMs, devido a sua alta acurácia;
- Potencial do uso desses métodos em áreas críticas, como saúde e finanças;
- Promessa de mais *Transparência* e *Segurança* para o uso de IA.

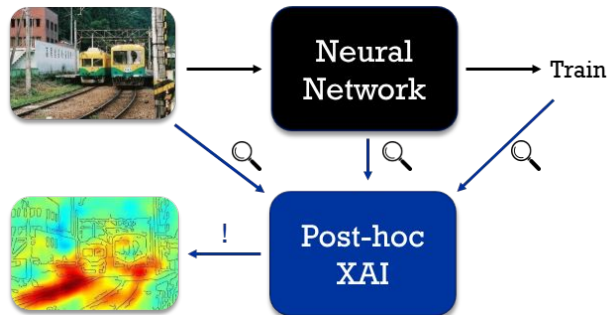
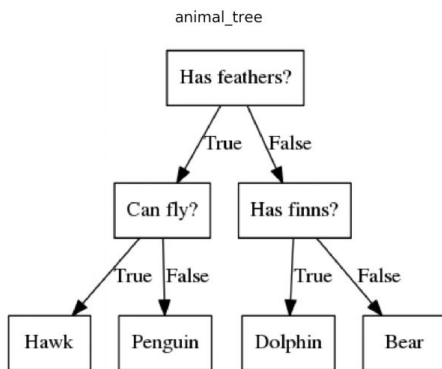


Número de artigos por ano com XAI como tema

eXplainable Artificial Intelligence

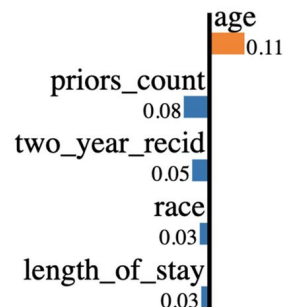
Métodos mais populares:

- Modelos de Interpretabilidade Intrínseca (modelos que podem ser interpretados por si só);
- Métodos post-hoc (buscam explicar os resultados obtidos pelos modelos), comumente utilizados com
- Ferramentas de Visualização (auxiliam na visualização do funcionamento do modelo).



High

Low





Trabalhos Recentes

Os trabalhos recentes em XAI buscam:

- Aumentar a *Transparência e Interpretabilidade* dos modelos de IA mais populares, como as LLMs;
- Possibilitar o uso de modelos de IA em áreas críticas, como saúde e finanças;
- Encontrar métodos *model-agnostic* e padronizados de XAI;
- Desenvolver medidas de performance para as XAI, de modo a possibilitar a comparação entre técnicas diferentes;
- Entre outras linhas de pesquisa.



Trabalhos Recentes - Saúde

Na área de saúde estão sendo utilizadas:

- As técnicas de SHAP e LIME para explicar previsões médicas e diagnósticos;
- O GradCAM para explicações em análises de imagem;
- Os modelos de Árvores de Decisão e Redes Bayesianas, devido a sua interpretabilidade intrínseca;
- Mecanismos de Atenção e Mapas de Saliência, para enfatizar regiões críticas em diagnósticos por imagem;
- Métricas de avaliação de XAIs, para fornecer uma base na decisão de qual modelo utilizar tendo em vista a criticidade e a velocidade necessária para determinada tarefa.



Trabalhos Recentes - Finanças

Na área de finanças o XAI, principalmente as técnicas SHAP e LIME, está sendo utilizado para:

- Detecção de Fraude não-invasiva;
- Análise de risco de crédito;
- Predição de mercado.



Trabalhos Recentes - Outras áreas

- **Educação:** Geração de explicações automáticas para ensino; Análise de dados vocacionais.
- **Cibersegurança:** Detecção de *malware* e inspeção de tráfego de rede.
- **Lei:** Interpretação de casos; análise de contratos; auxílio a decisões judiciais.
- **Agricultura:** Sistemas de recomendação de produto; análise de fatores que influenciam na qualidade dos produtos.



Proposta

- Relacionar Descrições e Imagens, utilizando o método de relacionamento de *embeddings* de textos e imagens;
 - Artigo base: Maniparambil et al. 2024: *Do vision and language encoders represent the world similarly?*
- Agrupar imagens e descrições, e relacionar a explicabilidade de ambos, utilizando algoritmos de XAI com os *embeddings* criados e as relações encontradas.



Experimentos

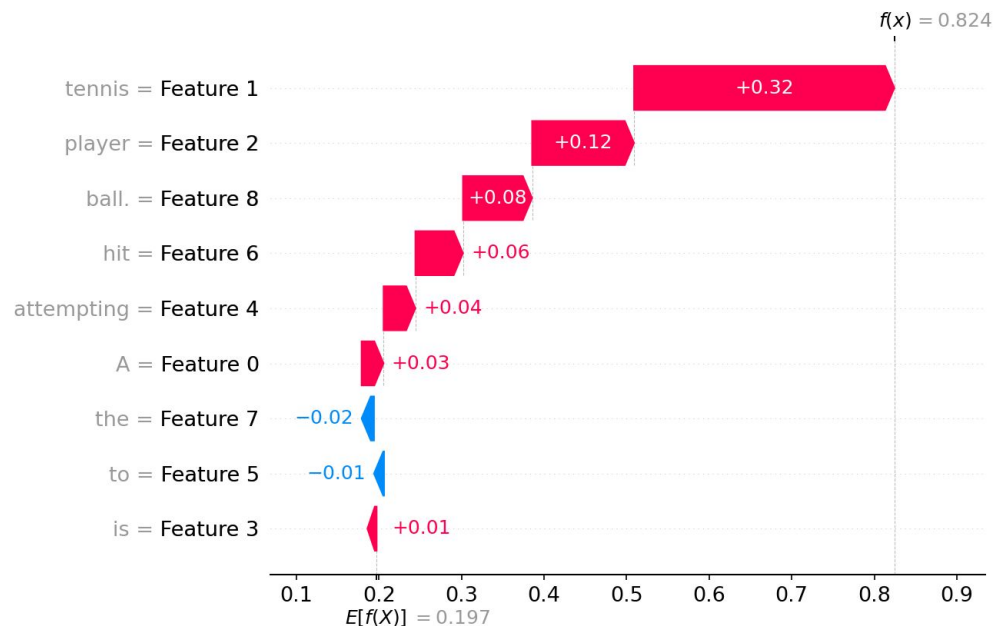
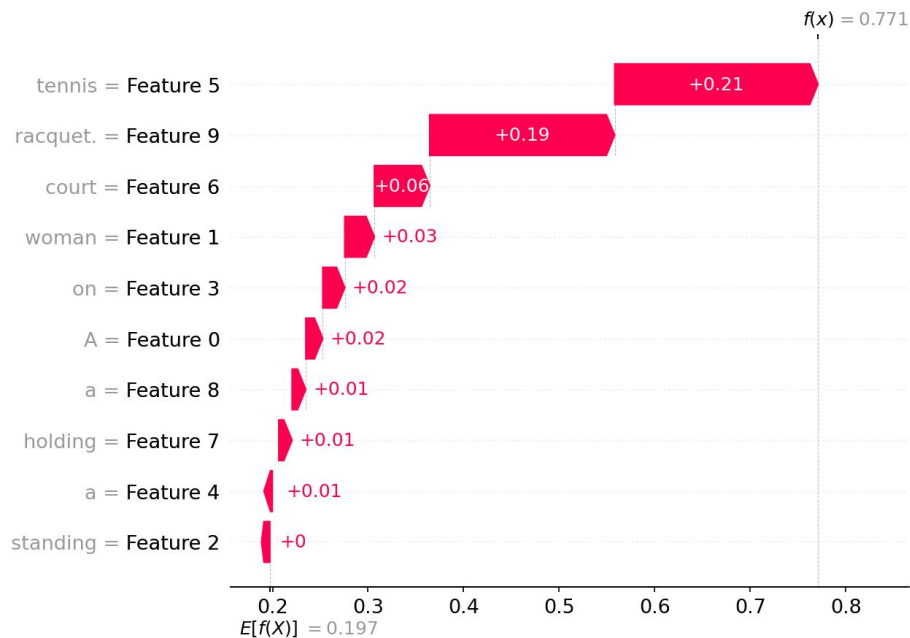
- Utilização do *COCO Dataset*;
- Para a criação dos *embeddings* de texto foi utilizado o modelo *RoBERTa*;
- Para a criação dos *embeddings* de imagem foi utilizado o modelo *ResNet50*;
- Para a explicabilidade do modelo de texto foi utilizado o algoritmo SHAP, utilizando as palavras como *features*;
- Para a explicabilidade do modelo de imagens foi utilizado o algoritmo Grad-CAM, diretamente em cima do resultado do *ResNet50*.

Resultados

A woman standing on a tennis court holding a racquet. / A tennis player is attempting to hit the ball.

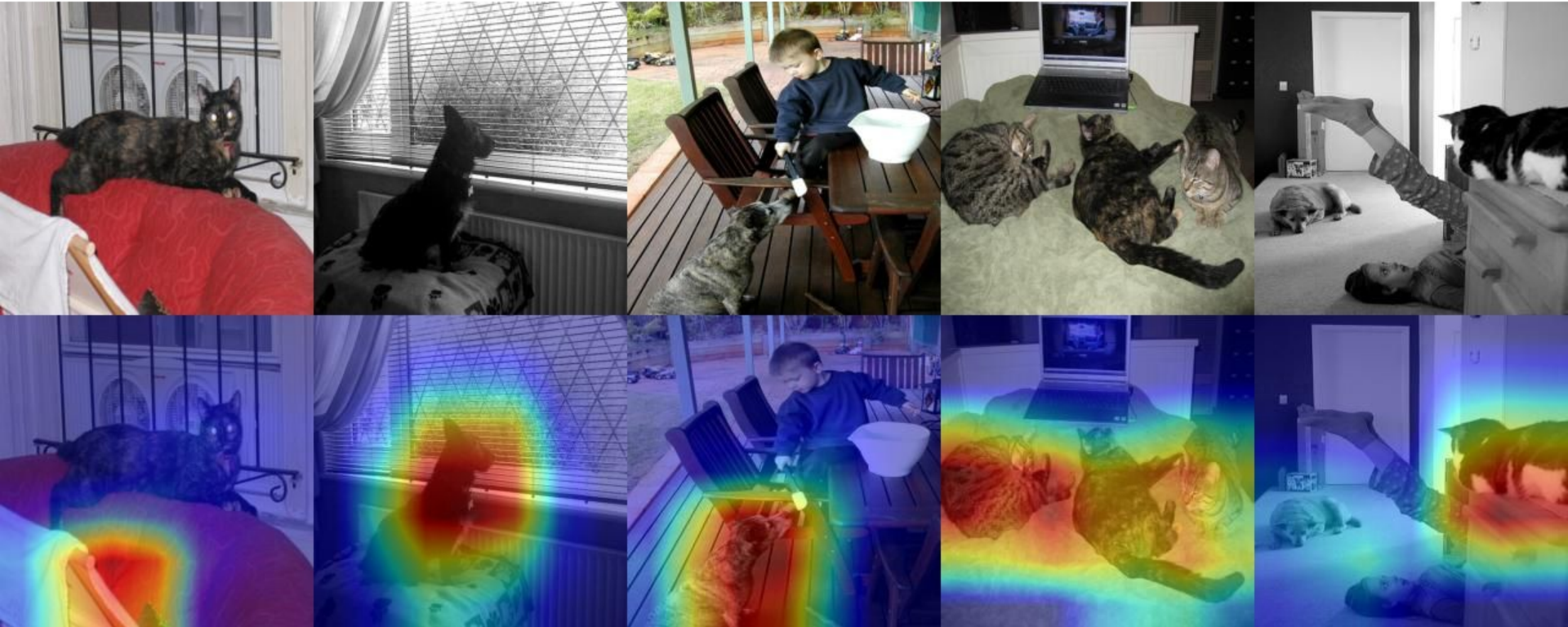


Resultados

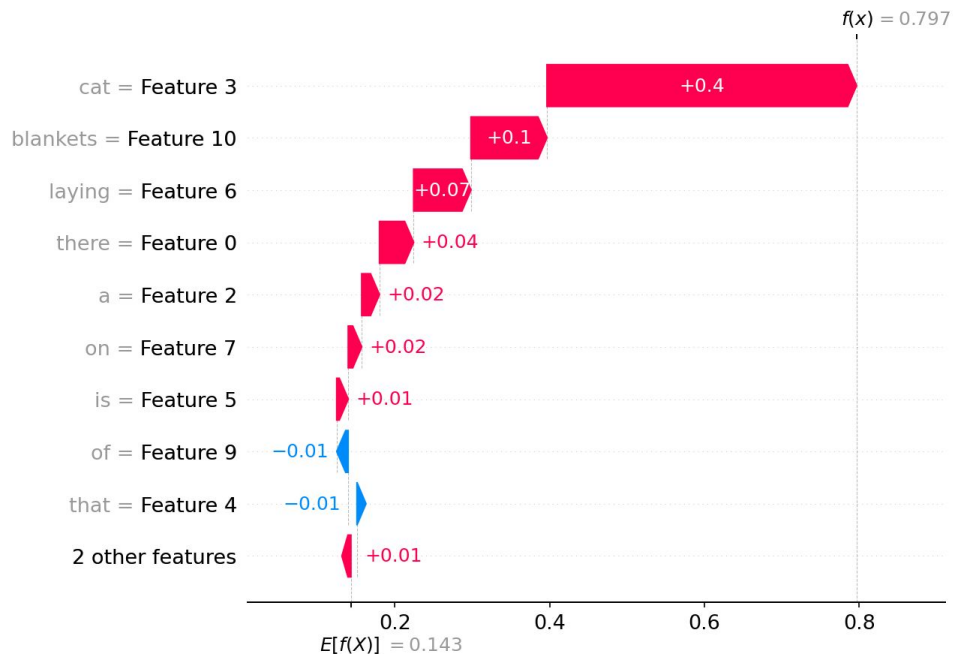
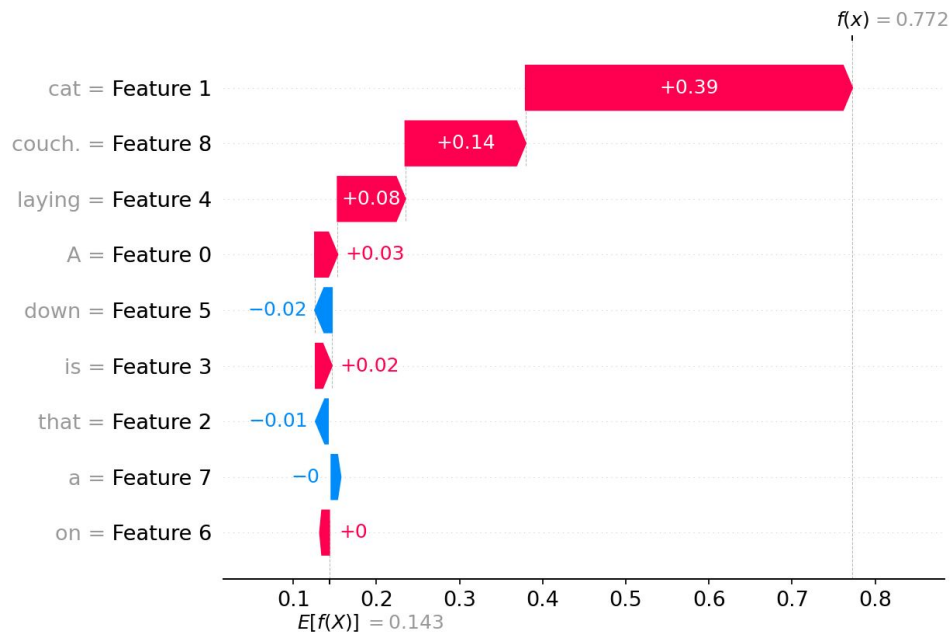


Resultados

A cat that is laying down on a couch. / there is a cat that is laying on top of blankets



Resultados

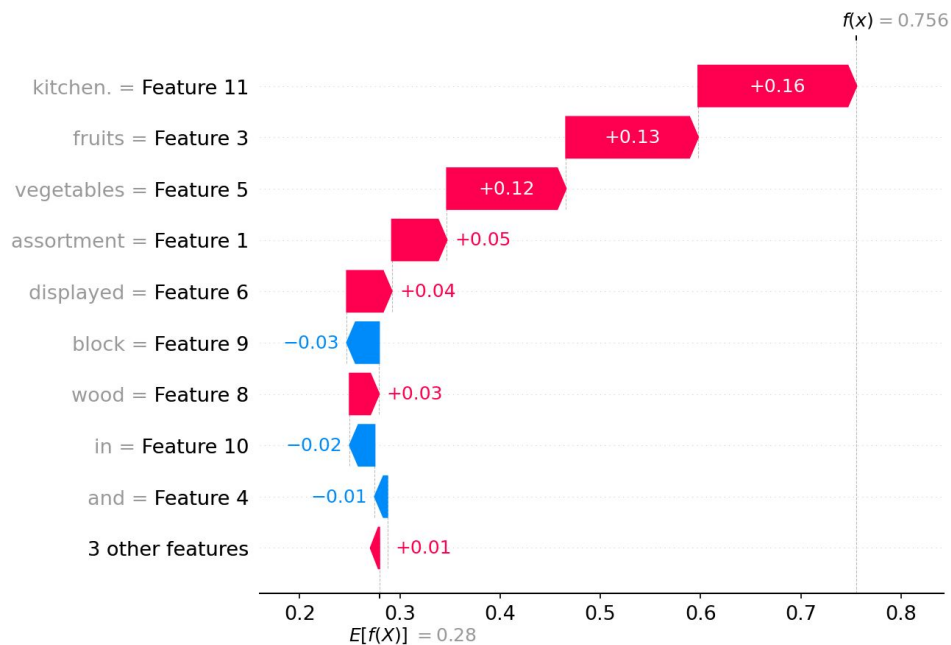
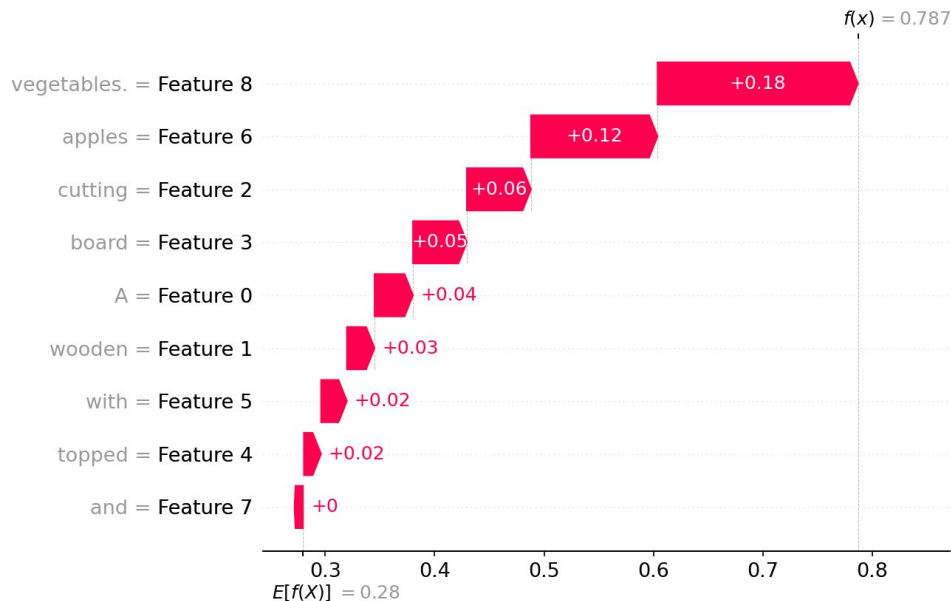


Resultados

A wooden cutting board topped with apples and vegetables. / Large assortment of fruits and vegetables displayed on wood block in kitchen.



Resultados





Conclusão

- O modelo do *ResNet50* possui alguns focos inesperados;
- Para a explicabilidade do modelo *RoBERTa*, é verificado um claro foco em substantivos e adjetivos, o que facilita a visualização das palavras importantes nas imagens;
- Poucas palavras colaboram de forma negativa para explicar seu resultado. Essas palavras são, em sua maioria, artigos e preposições.



Trabalhos Futuros - Experimentos

- Utilização de diferentes modelos para a criação dos *embeddings*;
- Utilização de diferentes algoritmos de XAI;
- Para a análise do *Grad-CAM*, utilizar a descrição relacionada, e não o resultado direto do modelo utilizado nos *embedds*;
- Realizar a análise completa em modelos generativos ou de busca de *image-to-text* ou *text-to-image*.



Trabalhos Futuros

Os trabalhos futuros de XAI envolvem:

- Criação de *benchmarks* e métricas para avaliação da performance das técnicas de XAI;
- Redução do *overhead* computacional relacionado aos métodos de XAI mais populares;
- Integração de *guidelines* éticos aos *frameworks* de XAI;
- Mapeamento das áreas críticas que exigem o uso de XAI na indústria;
- Mapeamento de itens das leis gerais de proteção de dados (LGPD, GDPR) que se alinham com o uso de XAI;
- Proposição de *frameworks* para o uso de XAI por leigos.

Obrigado!
