INTERPRETABILIDADE E CASUALIDADE

Trabalho Prático

Realizado por Carlota Santos, Nº51658
Docente: João Neves

Introdução

O presente trabalho aborda a interpretabilidade e causalidade dos modelos de *Machine Learning.* Com o intuito de obter informação fidedigna foi realizada a implementação e avaliação de métodos de interpretabilidade aplicados a um modelo de classificação de imagens.

A interpretabilidade dos modelos de *Machine Learning* é um aspeto fundamental para analisar e validar modelos complexos, especialmente redes neurais profundas usadas na classificação de imagens. Métodos de interpretabilidade *post-hoc* permite gerar explicações sobre as decisões dos modelos, normalmente sob a forma de mapas de saliência que destacam as regiões mais relevantes da imagem de entrada.

Por outro lado, a avaliação da qualidade dessas explicações não se deve limitar apenas à análise visual, sendo necessário recorrer a métricas quantitativas que permitam uma comparação objetiva e precisa entre diferentes métodos.

Neste trabalho prático, selecionamos algumas dessas métricas de interpretabilidade, baseadas na literatura científica, onde aplicamos essas métricas para avaliar diferentes técnicas de explicação num modelo de classificação de imagens.

O estudo utiliza diferentes técnicas de interpretabilidade e analisa os resultados através de métricas numéricas e gráficos. Dessa forma, é possível fazer uma comparação organizada e discutir de forma crítica tanto a utilidade quanto as limitações dos métodos e das métricas utilizados.

1. Artigos Selecionados e Fundamentação Teórica das Métricas

1.1 Artigo associado à métrica *Pointing Game*

O *Pointing Game* é uma métrica quantitativa utilizada para avaliar mapas de saliência, medindo a capacidade de um método de interpretabilidade, que pretende localizar corretamente as regiões relevantes da imagem associadas à decisão do modelo. Esta métrica é apresentada no artigo *Excitation Backprop for RNNs*, proposto por *Bargal et al.,* onde é utilizada para avaliar a localização espacial da evidência em modelos de visão computacional.

A motivação do *Pointing Game* reside na necessidade de avaliar explicações de forma objetiva, reduzindo a subjetividade inerente à análise visual. O funcionamento desta métrica consiste em identificar o ponto de maior ativação no mapa de saliência e verificar se este se encontra dentro da região anotada como relevante (*ground truth*). Se se verificar este pressuposto, regista-se um acerto; caso contrário, apura-se um erro, sendo a métrica final dada pela proporção de acertos.

Esta métrica apresenta inúmeras vantagens, de entre as quais se destacam a simplicidade, a fácil interpretação e a independência relativamente ao modelo ou método de interpretabilidade. No entanto, a métrica apresenta limitações, uma vez que considera apenas o ponto de maior ativação, ignorando a distribuição global da saliência, e também depende da existência de anotações de *ground truth*. Ainda assim, é particularmente adequada a problemas de classificação de imagens, onde a correção espacial da explicação é um critério relevante.

1.2 Artigo associado à métrica *Sparseness (Gini)*

A métrica de *Sparseness* é discutida no artigo *Concise Explanations of Neural Networks using Adversarial Training*, proposto por *Chalasani et al. (2020)*, no qual os autores defendem que explicações eficazes devem ser exatas. Nos mapas de saliência, esta exatidão é quantificada através do índice de Gini. Este é aplicado aos valores absolutos das atribuições, avaliando o grau de concentração da importância atribuída às diferentes regiões da imagem.

A *Sparseness* baseia-se na observação de que mapas de saliência, contudo se estes forem muito dispersos dificultam a interpretação humana. Pelo que se recorre ao índice de Gini, pois este assume valores entre 0 e 1, em que o zero representa uma distribuição uniforme da saliência, já os valores próximos de (1) indicam uma maior concentração da saliência em poucas regiões da imagem, tornando os mapas mais claros.

A métrica *de Sparseness* tem como vantagem, quantificar objetivamente a clareza visual das explicações. Mas, apresenta limitações importantes, uma vez que não avalia se as regiões destacadas são semanticamente ou espacialmente corretas, podendo atribuir valores elevados a explicações exatas, mas incorretas. Por este motivo, a *Sparseness* é mais eficaz quando utilizada em conjunto com métricas de correção espacial.

1.3 Justificação da escolha das métricas

As métricas selecionadas foram escolhidas por avaliarem dimensões complementares da qualidade das explicações. O *Pointing Game* permite analisar a correção espacial da explicação, verificando se a ativação máxima coincide com a região relevante da imagem. A *Sparseness*, baseada no índice de Gini, avalia a concentração da explicação, medindo o quão concisa é a distribuição da saliência.

A utilização conjunta destas métricas é relevante para o trabalho experimental realizado, uma vez que permite uma avaliação mais completa dos métodos de interpretabilidade, combinando correção espacial e clareza visual das explicações.

2. Metodologia

Neste capítulo descreve-se a metodologia seguida no trabalho experimental, apresentando o conjunto de dados utilizados, são estes o modelo de classificação e o pipeline.

2.1 Dataset utilizado

O trabalho experimental foi realizado utilizando o dataset **MNIST**, constituído por imagens de dígitos manuscritos em tons de cinzento. O dataset contém **10 classes**, correspondentes aos dígitos de 0 a 9. As imagens foram normalizadas de forma a garantir estabilidade durante o treino do modelo.

2.2 Modelo de classificação

Para a tarefa de classificação foi utilizada uma **rede neuronal convolucional (CNN)**, adequada a problemas de classificação de imagens. O modelo foi treinado de forma supervisionada do dataset MNIST. A definição da arquitetura e o processo de treino encontram-se implementados nos scripts model.py e train.py.

2.3 Pipeline experimental

O pipeline experimental seguido neste trabalho inclui o treino do modelo, a geração das explicações através dos métodos de interpretabilidade selecionados, o cálculo das métricas de avaliação e a agregação dos resultados. A execução global destas etapas é realizada pelo script main.py.

3. Métodos de Interpretabilidade Implementados

Nesta secção são descritos os métodos de interpretabilidade implementados e utilizados no trabalho experimental. Estes métodos permitem gerar explicações *post-hoc* para as previsões do modelo de classificação, sob a forma de mapas de saliência ou mapas de ativação, possibilitando a análise das regiões da imagem mais relevantes para a decisão do modelo.

3.1 Gradiente

O método do **Gradiente** consiste em calcular o gradiente da saída do modelo relativamente aos pixéis da imagem de entrada. Este gradiente indica a sensibilidade da previsão do modelo a pequenas variações em cada pixel, permitindo identificar quais as regiões da imagem que mais influenciam a decisão.

O tipo de explicação gerada corresponde a um mapa de saliência ao nível dos pixéis, onde valores mais elevados indicam maior influência na previsão.
A implementação deste método encontra-se no script explanations.py.

3.2 Integrated Gradients

O **Integrated Gradients** é uma extensão do método do gradiente simples que procura resolver problemas de interferência e instabilidade. O método calcula a média dos gradientes ao longo de

um caminho entre uma imagem de referência (baseline) e a imagem de entrada, integrando a contribuição de cada pixel.

Em comparação com o gradiente simples, o Integrated Gradients produz explicações mais estáveis e menos sensíveis a pequenas variações, oferecendo mapas de saliência mais consistentes.
Este método foi implementado no script explanations.py.

3.3 Occlusion

O método de **Occlusion** baseia-se na perturbação sistemática da imagem de entrada, ocultando pequenas regiões e observando o impacto dessa remoção na saída do modelo. A diminuição da confiança da previsão indica a relevância da região ocultada.

O tipo de perturbação aplicada consiste na substituição de regiões da imagem por um valor neutro, permitindo gerar um **mapa de importância baseado em perturbações**.
A implementação deste método encontra-se no script explanations.py.

3.4 Guided Backpropagation

O **Guided Backpropagation** é uma variação do método do gradiente que modifica a forma como os gradientes são propagados durante a retro propagação. Em particular, apenas gradientes positivos são considerados, tanto na passagem direta como na inversa.

O gradiente simples, é um método que tende a produzir mapas de saliência mais nítidos e focados, reduzindo a influência de regiões irrelevantes.
Este método foi implementado no script explanations.py.

3.5 Grad-CAM

O **Grad-CAM** é um método que utiliza os gradientes das camadas convolucionais finais do modelo para gerar mapas de ativação que indicam as regiões mais relevantes para uma determinada classe.

Ao contrário dos métodos baseados diretamente nos pixéis de entrada, o Grad-CAM produz explicações ao **nível das ativações internas do modelo**, resultando em mapas mais grossos, mas semanticamente mais interpretáveis.
A implementação deste método encontra-se no script explanations.py.

4. Métricas de Avaliação de Interpretabilidade

Aqui descrevem-se as métricas de interpretabilidade utilizadas na avaliação experimental dos métodos implementados. Ao contrário da Secção 2, onde foi apresentada a fundamentação teórica das métricas, aqui o foco incide na sua definição operacional, interpretação dos valores obtidos e aplicação prática no contexto experimental.

4.1 Pointing Game

O **Pointing Game** é utilizado para avaliar a correção espacial das explicações geradas pelos métodos de interpretabilidade. A métrica verifica se o ponto de maior ativação do mapa de saliência coincide com a região relevante da imagem definida pelo *ground truth*.

O valor do *Pointing Game* corresponde à proporção de amostras em que ocorre um acerto, assumindo valores entre 0 e 1. Valores mais elevados indicam uma maior capacidade do método em localizar corretamente a região relevante associada à decisão do modelo.

Para além da versão standard, foi utilizada a variante **Pointing Game Top-K**, na qual são considerados os *K* pontos de maior ativação do mapa de saliência. Um acerto é registado se pelo menos um desses pontos se encontrar dentro da região relevante, permitindo uma avaliação menos restritiva da explicação.

A implementação do *Pointing Game* e da sua variante Top-K encontra-se no script metrics_utils.py

4.2 Sparseness (Gini)

Conforme já foi referenciado anteriormente, e de forma mais sucinta a **Sparseness** é uma métrica utilizada para avaliar o grau de concentração dos mapas de saliência gerados pelos métodos de interpretabilidade. A métrica é calculada através do índice de Gini, aplicado aos valores absolutos do mapa de saliência.

Os valores da *Sparseness* situam-se no intervalo [0 , 1], sendo que valores mais elevados indicam explicações mais concentradas, enquanto valores mais baixos correspondem a mapas de saliência mais dispersos. Esta métrica está diretamente relacionada com mapas de saliência, uma vez que avalia a forma como a importância atribuída pelo modelo se distribui visualmente pela imagem.

A implementação da métrica de *Sparseness* encontra-se igualmente no script metrics_utils.py.

5. Resultados Experimentais

Aqui são apresentados os resultados obtidos com a aplicação das métricas de avaliação de interpretabilidade aos diferentes métodos considerados. Os resultados incluem valores quantitativos agregados, comparações gráficas entre métodos e exemplos de visualizações das explicações geradas.

5.1 Resultados quantitativos globais

Os resultados quantitativos globais foram obtidos através da aplicação das métricas *Pointing Game*, *Pointing Game Top-K* e *Sparseness* a um conjunto de amostras selecionadas do dataset. Os valores médios obtidos para cada método de interpretabilidade encontram-se resumidos no ficheiro results_summary.csv.

Estes resultados permitem uma comparação direta entre os diferentes métodos, evidenciando diferenças no desempenho ao nível da correção espacial das explicações e da concentração dos mapas de saliência.

5.2 Comparação entre métodos de interpretabilidade

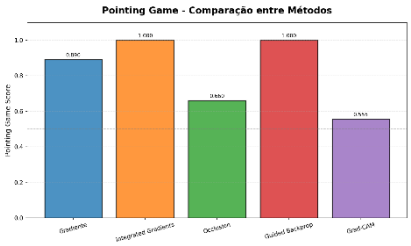
Para facilitar a análise comparativa, foram gerados vários gráficos que ilustram o desempenho dos métodos de interpretabilidade relativamente às métricas consideradas. Em particular, são apresentados gráficos comparativos para o *Pointing Game*, *Pointing Game Top-K* e *Sparseness*, permitindo observar de forma clara as diferenças entre métodos.

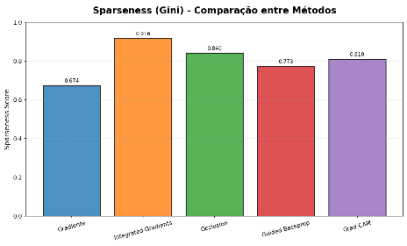
Estes gráficos fornecem uma visão global do comportamento dos métodos avaliados, servindo de base para a análise crítica apresentada na secção seguinte.

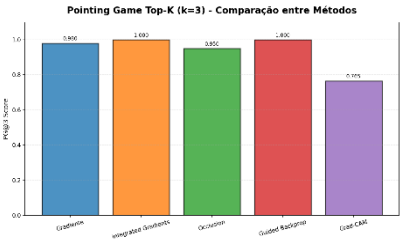
5.3 Visualizações das explicações

Para além dos resultados quantitativos, foram geradas visualizações das explicações produzidas pelos diferentes métodos para um conjunto de amostras representativas. Estas visualizações permitem uma análise qualitativa dos mapas de saliência e mapas de ativação, complementando os resultados obtidos pelas métricas quantitativas.

As visualizações ajudam a ilustrar as diferenças no tipo de explicação gerada por cada método e a relacionar os valores numéricos das métricas com padrões visuais observáveis.







A primeira figura representa a comparação dos métodos de interpretabilidade segundo a métrica Pointing Game.

Na segunda figura é possível observar a comparação dos métodos de interpretabilidade segundo a métrica Sparseness.

Na última figura observasse a comparação dos métodos de interpretabilidade segundo a métrica Pointing Game Top-k.

6. Análise Qualitativa das Explicações

Nesta secção é apresentada uma análise qualitativa das explicações geradas pelos diferentes métodos de interpretabilidade, recorrendo às visualizações produzidas para um conjunto de amostras selecionadas. O objetivo é complementar a análise quantitativa com uma inspeção visual dos mapas de saliência e mapas de ativação.

6.1 Visualizações das amostras selecionadas

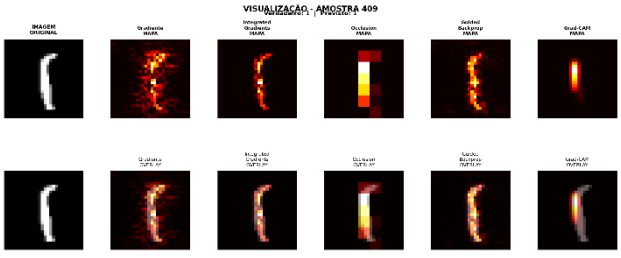
Foram geradas visualizações das explicações para um conjunto representativo de amostras do dataset, permitindo observar o comportamento dos diferentes métodos de interpretabilidade face à mesma imagem de entrada. Estas visualizações incluem mapas de saliência e mapas de ativação sobrepostos à imagem original.

As visualizações foram produzidas através do script visualize.py, que permite gerar e guardar automaticamente as explicações para cada método considerado.

6.2 Comparação visual entre métodos

A comparação visual entre os métodos evidencia diferenças claras no tipo de explicação gerada. Métodos baseados em gradientes tendem a produzir mapas mais detalhados ao nível dos pixéis, enquanto métodos como o Grad-CAM geram mapas mais suaves e localizados ao nível das ativações internas do modelo.

Estas diferenças refletem-se na clareza e concentração das explicações, sendo possível observar mapas mais focados em alguns métodos e mapas mais dispersos noutros, mesmo quando aplicados às mesmas amostras.



Exemplos de mapas de saliência e mapas de ativação gerados pelos diferentes métodos de interpretabilidade para amostras do dataset MNIST.

6.3 Relação preliminar entre mapas e métricas

A análise qualitativa permite estabelecer uma relação preliminar entre os mapas de saliência observados e os valores obtidos pelas métricas quantitativas. Em particular, mapas visualmente mais concentrados tendem a apresentar valores mais elevados de *Sparseness*, enquanto mapas que destacam corretamente a região relevante da imagem estão associados a melhores resultados no *Pointing Game*.

No entanto, esta análise também evidencia que explicações visualmente concisas nem sempre correspondem a explicações espacialmente corretas, reforçando a importância da utilização conjunta de múltiplas métricas na avaliação da interpretabilidade.

7. Discussão Crítica dos Resultados

Nesta secção é realizada uma análise crítica dos resultados obtidos, comparando os diferentes métodos de interpretabilidade avaliados com base nas métricas quantitativas e nas explicações visuais geradas. O objetivo é discutir as vantagens, limitações e adequação de cada método e métrica, evidenciando os principais trade-offs observados.

7.1 Comparação crítica entre métodos de interpretabilidade

A comparação entre os métodos avaliados revela diferenças significativas no tipo e na qualidade das explicações geradas. Métodos baseados em gradientes, como Gradiente simples, Integrated Gradients e Guided Backpropagation, tendem a produzir mapas de saliência mais detalhados ao nível dos pixéis, permitindo uma localização precisa das regiões relevantes da imagem.

O Integrated Gradients e o Guided Backpropagation destacam-se pela maior estabilidade e nitidez das explicações quando comparados com o gradiente simples, o que se reflete em melhores resultados nas métricas de correção espacial. Por outro lado, o Grad-CAM produz mapas mais suaves e de menor resolução espacial, mas semanticamente coerentes, ao focar-se nas ativações internas do modelo.

O método de Occlusion apresenta um comportamento distinto, uma vez que se baseia em perturbações diretas da imagem. Embora seja intuitivo e independente de gradientes, pode gerar mapas menos precisos espacialmente, dependendo do tamanho e da forma das regiões ocultadas.

7.2 Trade-offs entre métricas de avaliação

A análise conjunta das métricas evidencia trade-offs claros entre diferentes critérios de qualidade das explicações. O *Pointing Game* avalia exclusivamente a correção espacial da ativação máxima, enquanto a *Sparseness* avalia a concentração global da explicação.

Observa-se que métodos com bons resultados no *Pointing Game* nem sempre apresentam elevada *Sparseness*, indicando explicações corretas, mas potencialmente mais dispersas. Por outro lado, métodos com elevada *Sparseness* podem gerar explicações muito concentradas que nem sempre coincidem com a região semanticamente relevante da imagem.

A variante *Pointing Game Top-K* atenua parcialmente este trade-off, ao permitir que múltiplos pontos de elevada ativação sejam considerados, oferecendo uma avaliação menos restritiva da correção espacial.

7.3 Coerência entre métricas quantitativas e explicações visuais

De forma geral, existe uma boa coerência entre os resultados quantitativos e as explicações visuais observadas. Métodos que apresentam melhores valores no *Pointing Game* tendem a gerar mapas de saliência visualmente alinhados com as regiões relevantes da imagem, enquanto valores elevados de *Sparseness* estão associados a mapas mais concentrados e visualmente claros.

No entanto, também foram observados casos de incoerência, nos quais explicações visualmente concisas apresentam boa *Sparseness,* mas falham na correção espacial, resultando em valores mais baixos no *Pointing Game*. Estes casos demonstram que nenhuma métrica, isoladamente, é suficiente para avaliar completamente a qualidade das explicações.

7.4 Limitações das métricas utilizadas

Apesar da sua utilidade, as métricas utilizadas apresentam limitações importantes. O *Pointing Game* considera apenas o ponto (ou pontos) de maior ativação, ignorando a distribuição global da saliência e dependente da existência de anotações de *ground truth*. A *Sparseness*, por sua vez, não avalia a correção semântica da explicação, focando-se apenas na concentração dos valores.

Estas limitações reforçam a necessidade de utilizar múltiplas métricas complementares, bem como de combinar análise quantitativa com inspeção visual das explicações.

8.5 Adequação dos métodos e métricas a diferentes cenários

Os resultados indicam que a adequação dos métodos e métricas depende fortemente do contexto de aplicação. Em cenários onde a correção espacial é crítica, como tarefas de classificação de imagens com regiões bem definidas, o *Pointing Game* é particularmente relevante. Em contextos onde a clareza e concisão da explicação são prioritárias, a *Sparseness* constitui um critério útil.

Da mesma forma, métodos como Integrated Gradients e Guided Backpropagation são adequados quando se pretende uma explicação detalhada ao nível dos pixéis, enquanto o Grad-CAM é mais indicado para uma interpretação de alto nível baseada em regiões semanticamente relevantes.

Conclusão

Neste trabalho prático foi realizada a implementação e avaliação de métodos de interpretabilidade aplicados a um modelo de classificação de imagens, recorrendo a métricas quantitativas selecionadas a partir da literatura científica. Os resultados obtidos permitiram comparar diferentes abordagens de explicação, evidenciando diferenças relevantes tanto ao nível da correção espacial como da concentração das explicações geradas.

Os objetivos propostos foram alcançados, tendo sido implementadas métricas de avaliação de interpretabilidade e aplicadas a vários métodos de explicação. A análise quantitativa, complementada pela inspeção visual das explicações, permitiu uma avaliação comparativa rigorosa e sistemática dos métodos considerados.

Entre as principais aprendizagens destaca-se a importância da utilização conjunta de múltiplas métricas na avaliação da interpretabilidade. Os resultados demonstram que nenhuma métrica, isoladamente, é suficiente para capturar todas as dimensões da qualidade das explicações, sendo essencial combinar critérios de correção espacial e de clareza visual.

Como trabalho futuro, seria relevante aplicar as métricas a outros conjuntos de dados e arquiteturas de modelos, bem como explorar métricas adicionais de interpretabilidade. A integração de avaliações humanas e a análise do impacto das explicações na confiança dos utilizadores constituem igualmente direções interessantes para trabalhos futuros.

Referências

Bargal, S. A., Zisserman, A., & Hospedales, T. M. *Excitation Backprop for Recurrent Neural Networks*. European Conference on Computer Vision (ECCV).

Chalasani, P., Jha, S., Prasad, A., & Bastani, O. (2020). *Concise Explanations of Neural Networks using Adversarial Training*.