Mapa de calor GRAD-CAM como etapa em implementações de Visão Computacional

Bruno Luan Carvalho Leite Santos

[1]Departamento de Computação (DCOMP) – Universidade Federal de Sergipe (UFS)
Av. Marechal Rondon, s/n – Jardim Rosa Elze – CEP 49100-000
São Cristóvão – SE – Brazil

bruno.leite@dcomp.ufs.br

Resumo. Contexto: *Redes neurais são muitas vezes chamadas de "caixa-preta"devido a dificuldade de explicar um resultado em particular dado o elevado número de parâmetros treinados. Esse aspecto deve ser investigado, especialmente num contexto de inserção cada vez maior da visão computacional em cenários que envolvem riscos, como carros autônomos.* ***Objetivo:*** *Esse trabalho replica e avalia os resultados obtidos em trabalhos anteriores a respeito de transformar o mapa de calor GRAD-CAM numa etapa presente nos pipelines de visão computacional* ***Método:*** *Foi feita a replicação de experimento de trabalho anterior e seus resultados comparados.* ***Resultados:*** *Foram encontrados alguns padrões e feitas observações que não foram discutidas no artigo original, levando a necessidade de novas investigações sobre o modelo.* ***Conclusões:*** *a técnica de GRAD-CAM se mostrou válida para demonstrar o que está ativando a rede neural.*

1. Introdução

A visão computacional (CV) utilizando redes neurais convolucionais são cada vez mais presentes no apoio de atividades humanas que exijam analisar imagens e tomar decisões. Uma das aplicações onde essa tecnologia ganha mais espaço é como item de segurança em veículos [Falcini et al. 2017]. No entanto, pela própria natureza dessa aplicação - preservar a integridade dos ocupantes e os que estão ao alcance do veículo, o modelo de CV deve passar por avaliações quanto as áreas e pontos que estão alterando sua percepção da imagem e favorecendo alguma saída em detrimento das outras.

Um problema conhecido de modelos de aprendizado de máquina baseados em redes neurais é a difícil compreensão e consequente explicação dos motivos que levaram o modelo a classificar uma saída em detrimento das demais. Isso é rapidamente compreendido ao se ter em mente que os modelos podem atingir rapidamente milhões de parâmetros a serem treinados e posteriormente contribuírem para a saída do modelo. [Adadi and Berrada 2018], [Shen 2020].

Dada a crescente aplicação da Visão Computacional como item de segurança em diversos contextos - como câmeras de vigilância e veículos autônomos ou que contam com a tecnologia como auxiliar na segurança durante a condução, por exemplo, é crucial se entender ao máximo o que levou um modelo a se chegar a uma classificação específica de uma cena.

Com vistas a solucionar essa problemática, uma metodologia proposta foi observar visualmente quais áreas da imagem foram responsáveis por ativar a última camada

convolucional do modelo e visualizar no formato de mapa de calor sobreposto à imagem original [Falcini et al. 2017]. Um trabalho que utilizou essa metodologia para responder perguntas relacionadas a como o mapa de calor GRAD-CAM melhora o desenvolvimento de modelos de CV e sua aplicação como etapa em pipelines de aprendizado de máquina [Borg et al. 2021] é a base do presente trabalho. Assim, o objetivo geral deste trabalho é replicar o experimento realizado por Borg et al. (2021), afim de servir como validação daquele trabalho.

Este trabalho foi dividido em Introdução, onde foi contextualizado o mesmo; Fundamentação Teórica, que apresenta conceitos para o entendimento do trabalho; Metodologia, onde foram apresentadas as etapas de replicação do experimento realizado por Borg et al. (2021); Discussão dos resultados encontrados na replicação; e, finalmente, a Conclusão do trabalho.

2. Fundamentação Teórica

São abordados aqui conceitos de inteligência artificial, redes neurais convolucionais, GRAD-CAM e testagem de aprendizado de máquina. Além disso, uma breve explanação sobre confiabilidade da inteligência artificial no que tange a conformidade legal, ética e robustez técnica.

Aprendizado de máquina é termo bastante abrangente que se refere a uma grande variedade de algoritmos que possibilitam fazer predições, ou em outras palavras, inferências, baseados num conjunto de dados previamente conhecido. Comumente esses conjuntos de dados, ou datasets, são volumosos, chegando a milhões de registros únicos. Progressos no aprendizado de máquina possibilitam hoje atingir uma grande compreensão desses datasets, extraindo padrões entre os dados e informações que muitas vezes passam indetectáveis numa análise puramente humana. O aprendizado de máquina se tornou uma ferramenta poderosa, principalmente pelo grande número de dados disponíveis, crescimento exponencial do poder computacional e avanços em algoritmos da área.

Uma grande variedade destes algoritmos, chamados na área de modelos, estão sendo usados a todo momento em diversas áreas. A escolha de um determinado modelo é determinada tanto pelas características dos dados quanto pelo tipo de saída desejada. Grandes conjuntos de dados, da ordem de milhões de registros únicos, tendem a serem analisados por modelos mais sofisticados, como por exemplo algoritmos de aprendizado profundo. Conjuntos menores geralmente conseguem ser muito bem treinados e apresentar melhores resultados utilizando modelos mais simples, como técnicas baseadas em regressão ou até mesmo árvores de decisão, que cria subconjuntos a partir de regras aprendidas pelo modelo ao treinar com os dados apresentados. Reforça-se a importância de conhecer a natureza do domínio sendo estudado para se aplicar as técnicas mais adequadas, pois quando se fala de conjunto de dados, pode-se estar falando de imagens, áudios, registros temporais de uma variável independente, registros sociais, entre outros [Nichols et al. 2018].

Redes neurais artificiais são um tipo de modelo de aprendizado de máquina baseado no conceito de neurônios biológicos que consegue aprender padrões mesmo em conjuntos de dados extremamente grandes. Na realidade, quanto maior o conjunto, maior a capacidade da rede aprender e consequentemente passar a inferir a partir de novas observações. Da mesma forma que um conjunto de neurônios no cérebro humano, a

rede neural trabalha com camadas de neurôns ativando camadas subsequente até chegar na última camada, onde a rede irá produzir um resultado para as entradas que foram imputadas no início. As redes aprendem a partir da minimização do erro observado entre o resultado calculado por ela e o resultado real, fazendo ajustes nos parâmetros de cada neurôn da rede e até se chegar num número de rodadas de treinamento suficiente [Zhang et al. 2021].

As redes neurais convolucionais (RNN) são um tipo ainda mais específico. No contexto de imagens, este tipo de rede tem como entrada não somente um ponto ou pixel específico, mas sim uma região da imagem. O funcionamento dessa rede é aplicar operações chamadas de filtros para cada ponto dentro de uma região de uma imagem, enfatizando diferentes aspectos de cada vez que são passados adiantes na rede. Assim, região por região, a RNN varre toda a imagem, observando todas as partes que a compõe. Esse comportamento é importante para que o modelo de aprendizado não enxergue um pixel somente, mas sim o contexto onde o mesmo está inserido, observando a vizinhança de cada parte da imagem em relação as demais[Zhang et al. 2021].

Ao compreender o conceito básico do funcionamento das RNN, não é difícil entender que a saída da rede é afetada por um número extremamente grande de parâmetros. Em outras palavras, não fica claro que parte da imagem que serviu de entrada da rede contribuiu para uma determinada saída. Considerando que a saída da RNN de uma câmera pode ativar ações subsequentes, como frear um veículo ou fazer uma curva bruscamente, se torna essencial entender a motivação para ativação da rede em determinados contextos. Uma proposição para melhorar o entendimento dessas redes foi o GRAD-CAM. Dada uma categoria específica e uma imagem de entrada, o GRAD-CAM marca por meio de um mapa de calor as regiões da imagem que serviram de ativação para os neurôns da rede. O objetivo é garantir que o que a RNN viu na imagem (ou seja, o que ativou a rede) foi realmente o que deveria provocar o resultado de saída [Selvaraju et al. 2019]. Esse trabalho se trata da replicação realizada por Borg et al. (2021) como forma de avaliar os resultados encontrados por estes. Cabe ressaltar que o objetivo daquele trabalho foi adicionar a utilização do GRAD-CAM como parte de testes automatizados de modelos de aprendizado de máquina [Borg et al. 2021]

Um grupo denominado *High-Level Expert Group on AI (AI-HLEG)*, ou ”Grupo Expert de alto nível em inteligência artificial”, estabeleceu três componentes que devem ser observados em todo ciclo de vida de um modelo de aprendizado de máquina: 1 - Ter conformidade legal; 2 - Ser ético, garantindo aderência à princípios e valores; e 3 - Ter robustez, do ponto de vista técnico e humano, uma vez que pode causar danos não intencionais a terceiros. Esses conceitos em conjunto foram o que foi denominado como *Trustworthy AI*, ou ”Inteligência Artificial Confiável”[High-Level Expert Group on AI 2019].

3. Metodologia

A metodologia deste trabalho foi a replicação do experimento realizado por [Borg et al. 2021] e posterior comparação dos resultados encontrados. O papel da replicação de experimentos pode contribuir com o surgimento de interesse de novos pesquisadores. Alguns autores concordam com a hipótese de que a validação do procedimento seja um meio de estimular a propagação do conhecimento metodológico e uma

forma de assegurar a continuidade das investigações [Madden et al. 2013]. Portante, a importância da replicação reside, portanto, em garantir a validade e confiabilidade da correta condução metodológica, com o papel de proteger o leitor contra a aceitação de resultados não criticados [Singh et al. 2003].

4. Resultados e Discussão

O artigo original [Borg et al. 2021] cita que a acurácia do modelo atingiu 80%. Os autores citam que utilizaram a implementação open-source do GRAD-CAM demonstrada em [Chollet 2017] com incrementos realizados por aqueles autores. Alguns pontos chamam atenção do artigo original, em especial como há uma forte suspeita que a RNN não aprendeu o formato dos objetos, por exemplo. Em uma imagem houve uma classificação errada de cachorro sendo guiado por bicicleta. Os autores relacionaram isso ao fato de que em quase todas as imagens utilizadas no treino e possuíam cachorros, estes estavam na calçada. Nessa imagem em particular, o cachorro e seu condutor estavam na rua. Esse fato por si já desperta a necessidade de entender melhor o que as redes neurais estão de fato aprendendo do dataset. O código e todas as imagens com mapa de calor sobreposto podem ser encontrados no repositório criado para este artigo[1].

A réplica do experimento permitiu observar algo que não foi comentado no artigo original. Em alguns casos, houve uma ativação muito maior de diversas áreas da imagem aparentemente sem motivo, nem mesmo movimento havia nessas áreas. As figuras 1, 2 mostram a rede classificando uma bicicleta corretamente e um cachorro como bicicleta.

Figura 1. Classificação bicicleta correta

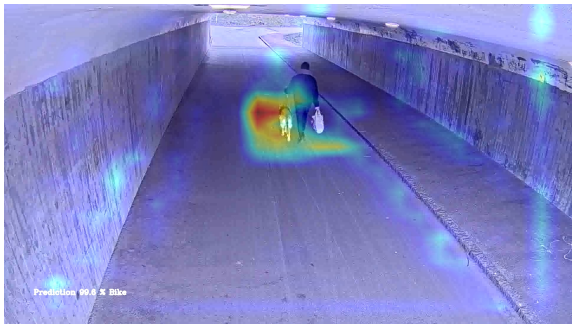

Figura 2. Classificação bicicleta errada

[1]https://github.com/brunoluanUFS/ETSN20

Nessas duas figuras, apesar de classificadas como bicicleta, quando a imagem tem um cachorro fica nítido que mais regiões foram atividas, porém talvez não o suficiente para alterar uma eventual classificação para cachorro. Ademais, cabe destacar que o cachorro estava andando pela rua, o que pode ter confundido o modelo, conforme já apontado pelos autores do artigo original.

Já nas figuras 3, 4 e 5, a classificação foi feita corretamente como cachorro, porém o mapa de calor ficou muito mais ativado nas mesmas regiões da imagem, mesmo sem nenhum objeto diferente nestas regiões.

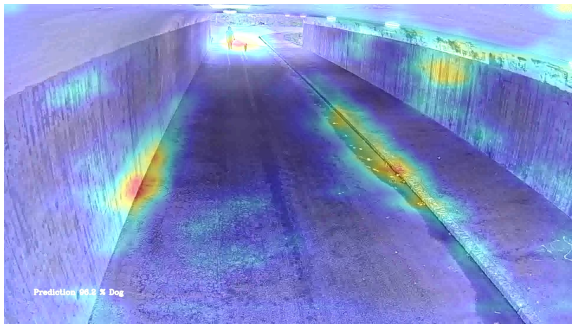

Figura 3. Mapa de calor ativado em regiões sem atividade

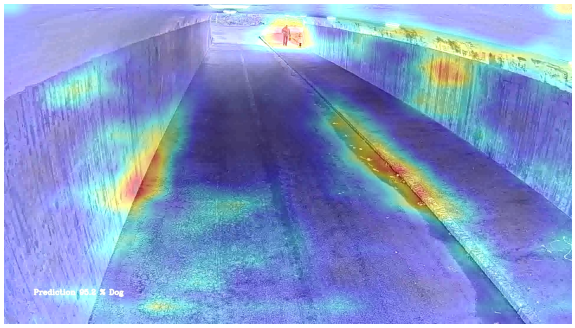

Figura 4. Mapa de calor ativado em regiões sem atividade

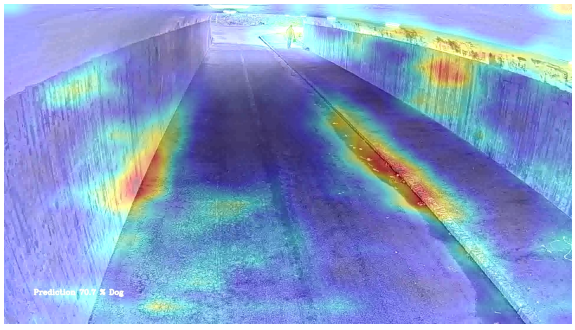

Figura 5. Mapa de calor ativado em regiões sem atividade

Uma possibilidade pode ser que as imagens que foram disponibilizadas para treinamento possuíam algum tipo de efeito visual, como iluminação por exemplo, que alteraram os pesos do modelo para essa categoria específica. É necessário conduzir um estudo mais aprofundado sobre esse ponto.

5. Conclusão

Neste trabalho foi feita a replicação do experimento de classificação de imagens utilizando uma rede neural previamente treinada. O código utilizado foi encontrado no artigo de [Borg et al. 2021].

Foram encontrados algumas particularidades observados nas imagens classificadas que não foram discutidas no artigo original. Porém, se trata de uma investigação a ser realizada no processo de treinamento e parametrização do modelo da RNN. Como técnica de apoio, o GRAD-CAM demonstrou validade ao apontar visualmente as partes da imagem que estão ativando a rede e resultando em determinadas classificações, o que é o objetivo da ferramenta.

Como atividade futura, sugere-se investigar os motivos que levaram algumas imagens a serem ativadas em regiões onde não havia nenhum tipo de movimento. Também sugere-se investigar se o modelo realmente está ponderando muito a região onde o objeto aparece na imagem, por exemplo bicicleta na rua e cachorro na calçada.

Referências

[Adadi and Berrada 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[Borg et al. 2021] Borg, M., Jabangwe, R., Åberg, S., Ekblom, A., Hedlund, L., and Lidfeldt, A. (2021). Test automation with grad-cam heatmaps - A future pipe segment in mlops for vision ai? *CoRR*, abs/2103.01837.

[Chollet 2017] Chollet, F. (2017). 5.4-visualizing-what-convnets-learn.

[Falcini et al. 2017] Falcini, F., Lami, G., and Costanza, A. M. (2017). Deep learning in automotive software. *IEEE Software*, 34:56–63.

[High-Level Expert Group on AI 2019] High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy ai. Report, European Commission, Brussels.

[Madden et al. 2013] Madden, C., Easley, R., and Dunn, M. (2013). How journal editors view replication research. *Journal of Advertising*, 24:77–87.

[Nichols et al. 2018] Nichols, J., Chan, H., and Baker, M. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, 11.

[Selvaraju et al. 2019] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

[Shen 2020] Shen, O. (2020). Interpretability in ml: A broad overview. *The Gradient*.

[Singh et al. 2003] Singh, K., Ang, S. H., and Leong, S. (2003). Increasing replication for knowledge accumulation in strategy research. *Journal of Management*, 29:533–549.

[Zhang et al. 2021] Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning.