Análise do Método de Extração de Atributos Visuais através de Camadas Ocultas de uma ConvNet Residual

Analysis of the Visual Attribute Extraction Method through Hidden Layers of a Residual ConvNet

Vinícius Cerqueira Ribeiro[1], Thiago França Naves[2], Arlete Teresinha Beuren[3]

RESUMO

Este estudo visa aprofundar a compreensão e análise do método de extração de características por meio de camadas ocultas em uma rede convolucional residual, conforme proposto por Baloian, Murrugarra-Llerena e Saavedra (2021) em seu artigo "Scalable Visual Attribute Extraction through Hidden Layers of a Residual ConvNet". Utilizando o mesmo conjunto de dados criado pelos autores, que inclui grupos de cor e textura, conduzimos testes com a ResNet-50. Além disso, exploramos a capacidade de generalização dos resultados por meio da aplicação da rede VGG-16. Nossos resultados confirmaram que, para esses conjuntos de dados, as camadas 2 e 4 da ResNet-50 apresentaram os melhores desempenhos, corroborando as descobertas dos autores originais. Além disso, observamos a mesma tendência em direção às camadas mais superficiais da VGG-16 para a extração de características relacionadas à cor e às camadas mais profundas para características de textura, neste caso as camadas 1 e 4. A replicação bem-sucedida desses resultados valida a confiabilidade desses métodos na extração de atributos visuais sem a necessidade de treinamento extensivo, destacando seu amplo potencial de aplicação na visão computacional.
PALAVRAS-CHAVE: Extração de Características; Redes Convolucionais; Comércio Eletrônico.

ABSTRACT

This study aims to deepen the understanding and analysis of the method for feature extraction through hidden layers in a residual convolutional network, as proposed by Baloian, Murrugarra-Llerena, and Saavedra (2021) in their paper "Scalable Visual Attribute Extraction through Hidden Layers of a Residual ConvNet.". Utilizing the same dataset created by the authors, which includes groups of color and texture, we conducted tests with the ResNet-50. Additionally, we explored the generalization capability of the results by applying the VGG-16 network. Our results confirmed that, for these datasets, layers 2 and 4 of the ResNet-50 exhibited the best performance, corroborating the original authors' findings. Furthermore, we observed the same trend towards the shallower layers of the VGG-16 for color-related feature extraction, and the deeper layers for texture-related features, specifically layers 1 and 4 in this case. The successful replication of these results validates the reliability of these methods in visual attribute extraction without the need for extensive training, highlighting their broad potential application in computer vision.
KEYWORDS: Feature Extraction; Convolutional Networks; eComerce.

[1] Bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico. Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil. E-mail: viniciuscerqueira@alunos.utfpr.edu.br. ID Lattes: 0815282128812756.
[2] Docente no Curso de Ciência de Computação. Universidade Tecnológica Federal do Paraná, Santa Helena, Paraná, Brasil. E-mail: naves@utfpr.edu.br ID Lattes: 2177644773849043.
[3] Docente no Curso de Ciência da Computação. Universidade Técnológica Federal do Paraná, Santa Helena, Paraná, Brasil. E-mail: arletebeuren@utfpr.edu.br. ID Lattes: 0084145280240578.

INTRODUÇÃO

Os avanços na área de redes convolucionais têm impulsionado o desenvolvimento de soluções dinâmicas e robustas para a industria, especialmente no âmbito da visão computacional. Um exemplo dessas aplicações está relacionada com a recuperações de imagens no comércio eletrônico, por meio de texto, imagem ou uma união de ambos. Cada vez mais, surgem demandas por soluções inovadoras para atender a essa necessidade crescente.

Hoje a busca por texto ainda prevalece como a principal modalidade de consulta adotada pelos motores de busca, no entanto sua eficácia depende diretamente de uma descrição detalhada do produto. Em alguns casos, para aprimorar a experiência de busca, os motores incorporam a funcionalidade de busca por imagens, trazendo resultados muito mais precisos (DUBEY, 2021).

Sendo assim, este trabalho busca analisar o método de extração de características utilizando redes convolucionais pré-treinadas apresentado por Baloian, Murrugarra-Llerena e Saavedra (2021) e repetir seus resultados bem como verificar a replicabilidade dos experimentos para outras redes. O objetivo é compreender como diferentes camadas de uma mesma rede se comportam na extração de atributos visuais de uma imagem.

O trabalho está estruturado em três seções principais: "Materiais e Métodos", que descreve a abordagem adotada para os experimentos; "Resultados e Discussões", onde são apresentados os resultados da aplicação da metodologia e suas implicações; e, por fim, as conclusões.

MATERIAIS E MÉTODOS

Para a primeira etapa deste estudo foi necessário a análise e revisão dos testes propostos no artigo Scalable Visual Atribute Extraction through Hidden Layers of a Residual ConvNet, escrita por Baloian, Murrugarra-Llerena e Saavedra (2021), com o objetivo de compreender e replicar os resultados e observações sobre a extração de características dos blocos residuais de uma rede ResNet-50 com seus pesos pré-treinados.

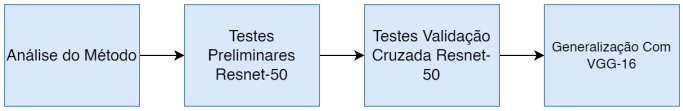
De acordo com Baloian, Murrugarra-Llerena e Saavedra (2021), para a construção do cojunto de dados para teste foi utilizada a plataforma Kaggle e outras fontes online. Dois conjuntos de dados foram elaborados, o primeiro categoriza suas imagens por Cor seguindo as seguintes classes: vermelho, preto, azul, verde, amarelo, cinza, marrom, rosa, roxo, e laranja. O segundo, categoriza suas imagens por Textura com as seguintes classes: xadrez, listrado, florida, leopardo, poá, básico, caxemira, argyle, pé de ganso, lantejoula. Cada classe contém 100 imagens, totalizando 1000 imagens em cada *dataset*. O conjunto foi disponibilizado pelos autores e é o mesmo que foi utilizado para este estudo.

A extração de características foi realizada utilizando a rede pré-treinada no conjunto ImageNet, ResNet-50, introduzida por He *et al.* (2016). Foi renomeado a sua primeira camada convolucional para Bloco 1, e os demais blocos residuais da rede foram renomeados para Bloco 2 ao Bloco 5, conforme descrito no artigo original. Na saída de cada um destes blocos, foi aplicada uma camada de *Global Avarage Pooling 2D* (GAP) para redução do espaço dimensional, e as saídas dessa camadas forma utilizadas como entrada para um classificador *K-Neares Neighbour* (KNN). Para a validação do

modelo, foi utilizada a técnica de validação curzada, ou *cross-validadtion*, com *5-fold*, o mesmo valor indicado pelos autores.

Além dos testes conduzidos com a ResNet-50, também foram realizados experimentos utilizando a rede VGG-16, uma rede convolucional profunda clássica apresentada por Simonyan e Zisserman (2014). A rede também possui uma versão pré-treinada no conjunto de dados da ImageNet. O objetivo desses testes foi generalizar os resultados obtidos com a ResNet-50 para a VGG-16, a fim de verificar se os comportamentos das camadas são replicados. De acordo com Li *et al.* (2015) redes treinadas em um mesmo problema podem aprender características básicas similares em uma camada ou conjunto de camadas, como cor, silhuetas e texturas.

Figura 1 – Etapas para obtenção dos resultados.



Fonte: Autoria Própria (2023).

RESULTADOS E DISCUSSÕES

Nesta seção, serão apresentados e discutidos os resultados e observações obtidas durante a realização dos experimentos descritos na seção anterior. Inicialmente, serão apresentados os resultados de Baloian, Murrugarra-Llerena e Saavedra (2021). Em seguida, abordaremos a replicação da metodologia desses autores, utilizando tanto a ResNet-50 quanto a rede VGG-16, e discutiremos os resultados obtidos.

Conforme a avaliação de Baloian, Murrugarra-Llerena e Saavedra (2021), foi constatado que, em relação à característica de cor, o segundo bloco da rede ResNet-50 demonstrou uma acurácia mais elevada, bem como uma melhor separação dos grupos quando utilizado a visualização do espaço com a técnica de *Uniform Manifold Approximation and Projection* (UMAP) (MCINNES; HEALY; MELVILLE, 2018). Por outro lado, em relação à textura, o quarto bloco exibiu melhores resultados. Esta análise corrobora a observação de que as redes neurais tendem a adquirir compreensões de conceitos de baixo nível nas camadas iniciais (como no caso da cor), enquanto conceitos intermediários são mais proeminentes em camadas mais profundas da rede (como a textura). (RAFEGAS et al., 2020).

Tabela 1 – Resultados para as 5 camadas ResNet-50.

#Bloco	Cor	Textura
Bloco 1	90,0	58,7
Bloco 2	**93,7**	82,6
Bloco 3	92,9	91,5
Bloco 4	88,6	**93,9**
Bloco 5	78,4	92,5

Fonte: Baloian, Murrugarra-Llerena e Saavedra (2021).

Os resultados expostos nas Tabelas 2 e 3 seguem a mesma metodologia empregada pelos autores no artigo, conforme previamente detalhado na seção anterior. A Tabela 2 exibe a acurácia de

cada um dos blocos, com o conjunto de dados dividido em 70% para treino e 30% para validação (um teste preliminar de verificação). Por sua vez, a Tabela 3 apresenta a acurácia obtida através de uma validação cruzada com 5 conjuntos (*5-folds*), proporcionando uma análise mais abrangente e robusta das métricas de desempenho.

Tabela 2 – Resultados para as 5 camadas ResNet-50 (70/30).

#Bloco	Cor	Textura
Bloco 1	87,0	27,0
Bloco 2	**91,7**	79,7
Bloco 3	91,3	**93,0**
Bloco 4	87,0	91,7
Bloco 5	64,3	87,7

Fonte: Autoria Própria (2023).

Os resultados apresentados revelam uma discrepância em relação ao estudo original. Diferentemente da Tabela 1, o experimento indica que a terceira camada apresenta uma melhor acurácia para extração de atributos de textura. Contudo, vale ressaltar que a divisão aleatória do conjunto de treino e validação, principalmente em conjuntos de dados pequenos,pode resultar na seleção de subgrupos subótimos para validação ou treinamento.

A utilização da validação cruzada aborda essa questão de maneira eficaz. Nesse método, o conjunto de dados completo é aleatorizado e dividido em K subconjuntos. Em cada iteração em K, um dos subconjuntos é isolado para validação, enquanto os demais são destinados ao treinamento, ao final a resultado do modelo em cada uma das iterações é combinada. Essa abordagem sistemática reduz a influência de uma única divisão aleatória e proporciona uma avaliação mais robusta e confiável do desempenho do modelo(KOHAVI et al., 1995).

Tabela 3 – Resultados para as 5 camadas ResNet-50 (cross-validation).

#Bloco	Cor	Textura
Bloco 1	88,4	29,3
Bloco 2	**94,2**	82,1
Bloco 3	93,9	91,5
Bloco 4	85,8	**94,4**
Bloco 5	70,9	90,6

Fonte: Autoria Própria (2023).

Ao empregar a validação cruzada, observa-se que as camadas 2 e 4 demonstram um desempenho superior para cor e textura, respectivamente. Esses resultados exibem uma proximidade com os dados destacados na Tabela 1, validando e reforçando as conclusões dos autores quanto à habilidade das redes convolucionais residuais em aprender atributos distintos em diversos níveis de sua arquitetura.

Na sequência, foram realizados os testes com a mesma base de imagens usando a rede convolucional VGG-16, seguindo a mesma metodologia para extração de características.

Tabela 4 – Resultados para as 5 camadas VGG-16 (cross-validation).

#Bloco	Cor	Textura
Bloco 1	**92,2**	47,4
Bloco 2	88,8	76,1
Bloco 3	88,1	89,2
Bloco 4	80,9	**93,9**
Bloco 5	57,1	85,4

Fonte: Autoria Própria (2023).

Os resultados obtidos com a arquitetura VGG-16 corroboraram a tendência de melhor desempenho na extração de características relacionadas a cores nas camadas iniciais e, na quarta camada, para características relacionadas à textura. No entanto, é importante notar que, mesmo seguindo essa tendência, os resultados da VGG-16 não conseguiram superar os valores alcançados pela arquitetura ResNet-50 no mesmo conjunto de dados. Entre as duas arquiteturas analisadas, a ResNet-50 se destacou como mais eficaz na extração de características para o problema em questão. Notavelmente, as camadas 2 e 4 da ResNet-50 demonstraram um desempenho superior na extração de características de cor e textura, respectivamente.

CONCLUSÕES

Em suma, a replicação bem-sucedida dos resultados do estudo anterior representa um passo significativo na consolidação das descobertas relacionadas à extração de atributos visuais específicos através das camadas de redes convolucionais pré-treinadas. A validação dessas observações quanto à capacidade de identificar características específicas e classificar categorias de forma eficaz, sem a necessidade de treinamento exaustivo da rede, reforça a confiabilidade e a utilidade desses métodos no campo da visão computacional em especial no contexto dos motores de busca por imagens, já que permite muito mais facilmente a escalabilidade em comparação com redes convolucionais treinadas para classificação. Além disso, a capacidade de generalizar esses resultados para outras arquiteturas, como a VGG-16, destaca a versatilidade e a robustez das conclusões alcançadas, o que amplia seu potencial de aplicação.

Agradecimentos

Expressamos nossa sincera gratidão às entidades de fomento à pesquisa, em especial à Universidade Tecnológica Federal do Paraná (UTFPR), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação Araucária, pelo apoio fundamental e recursos disponibilizados para a realização deste trabalho de pesquisa. Suas contribuições foram essenciais para o sucesso deste estudo.

Conflito de interesse

Não há conflitos de interesses.

REFERÊNCIAS

BALOIAN, Andres; MURRUGARRA-LLERENA, Nils; SAAVEDRA, Jose M. Scalable visual attribute extraction through hidden layers of a residual convnet. **arXiv preprint arXiv:2104.00161**, 2021.

DUBEY, Shiv Ram. A decade survey of content based image retrieval using deep learning. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 32, n. 5, p. 2687–2704, 2021.

HE, Kaiming et al. Deep Residual Learning for Image Recognition. In: 2016. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: [s.n.], 2016. P. 770–778. DOI: 10.1109/CVPR.2016.90.

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA, 2. IJCAI. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

LI, Yixuan et al. Convergent learning: Do different neural networks learn the same representations? **arXiv preprint arXiv:1511.07543**, 2015.

MCINNES, Leland; HEALY, John; MELVILLE, James. Umap: Uniform manifold approximation and projection for dimension reduction. **arXiv preprint arXiv:1802.03426**, 2018.

RAFEGAS, Ivet et al. Understanding trained CNNs by indexing neuron selectivity. **Pattern Recognition Letters**, Elsevier, v. 136, p. 318–325, 2020.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.