Uso de deep learning na classificação de imagens de produtos de um pet shop

Vinícius D. A. Figueiredo[1], Thiago M. Ventura[1], Raphael de S. R. Gomes[1], Fabio S. Vitoriano[1].

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT), 78060-900 Cuiabá – MT

viniciusdotto@hotmail.com, thiago@ic.ufmt.br, raphael@ic.ufmt.br, fabio.vitoriano@ufmt.br

Abstract. *Product recognition is a valuable tool for applications aimed at promoting accessibility and process automation. The objective of this work was to classify products specifically from a pet shop through image recognition using deep learning techniques. For training the model, it was necessary to build a custom image dataset. In total, 100 classes were selected, covering the main products. Each class was populated with thousands of examples to allow the model to abstract the key characteristics of each product. After training the model using convolutional neural networks, the model achieved an average accuracy of 88.21%, with the best result being 89.77%.*

Resumo. *O Reconhecimento de produtos é uma ferramenta útil para aplicações voltadas à promoção de acessibilidade e à automação de processos. O objetivo deste trabalho foi conseguir classificar produtos especificamente de um pet shop, através do reconhecimento de imagens utilizando técnicas de deep learning. Para o treinamento do modelo, foi necessária a construção de uma base de imagens própria. No total, 100 classes foram selecionadas, abrangendo os principais produtos. Cada classe foi alimentada com milhares de exemplos para que fosse possível ao modelo abstrair as principais características de cada produto. Após treinamento do modelo utilizando redes neurais convolucionais, o modelo obteve uma acurácia média de 88,21%, alcançando o melhor resultado de 89,77%.*

1. Introdução

O reconhecimento automático de produtos em ambientes de varejo oferece diversos benefícios, especialmente ao melhorar a acessibilidade e a experiência de compra dos consumidores. Por exemplo, em Merler et al. (2007), foi demonstrado que o uso de técnicas de reconhecimento de imagem para identificar produtos de mercearia pode ser uma ferramenta valiosa para auxiliar pessoas com deficiência visual, no qual os produtos são identificados e as suas informações transmitidas por áudio. Trabalhos como os de Tonioni et al. (2018) e Franco et al. (2017) expandiram essa visão ao explorar o potencial de aplicações interativas com realidade aumentada, melhorando significativamente a experiência do consumidor no ponto de venda e possibilitando uma interação mais rica com os produtos disponíveis.

O reconhecimento de objetos no contexto da visão computacional, descrito por Szeliski (2010) evoluiu ao longo do tempo, especialmente com o avanço das técnicas de aprendizado de máquina e, mais especificamente, de deep learning. Essas técnicas trouxeram precisão e eficiência inéditas para a identificação automática de objetos, permitindo sua aplicação em uma variedade de setores, incluindo o varejo. Mais recentemente, conforme mencionado por Selvam et al. (2024), o setor de varejo tem acompanhado esses avanços tecnológicos, adaptando-se para atender às demandas de mercados especializados, como o de produtos para animais de estimação.

Neste contexto, o objetivo deste trabalho é desenvolver um modelo de deep learning voltado para a classificação de produtos na área de pet shop. Esse modelo pretende otimizar a gestão e o reconhecimento desses produtos, criando uma solução automatizada que agilize processos operacionais, reduza erros e melhore o atendimento ao cliente.

2. Materiais e Métodos

2.1. Classes e geração do banco de imagens

Para o treinamento do modelo é necessário um banco de imagens robusto, contendo imagens nos mais variados ângulos e iluminações. De acordo com Pietrini (2024), existem vários *datasets* de imagens disponíveis, tais como SOIL-47 por Koubaroulis et al. (2002), Grozi-120 por Merler et al. (2007). Entretanto, as bases citadas são de uso do varejo em geral não sendo especializado em pet shops, sendo inadequadas para treinamento do modelo proposto. Portanto, foi necessária à construção de uma base de imagens própria para pet shop.

Na construção da base de imagens, vídeos de cada produto foram gravados para posterior geração de diversas imagens. As gravações foram realizadas em diferentes distâncias, com os produtos posicionados em vários ângulos e ambientes distintos, garantindo variabilidade nas imagens. Desta forma, cada produto foi capturado com planos de fundo e condições de iluminação variadas — incluindo luz natural, artificial e pouca iluminação. A Figura 1 ilustra alguns exemplos da obtenção de imagens de diferentes ângulos e diferentes condições de iluminação.

Figura 1. Exemplo de angulações, iluminações e plano de fundo.

Após a captura dos vídeos, foi realizada a conversão do vídeo em imagens. Para a conversão foi utilizado o software gratuito *Free Video to JPG Converter*, sendo gerado 65.412 imagens dentre os produtos. Todas as imagens estão no formato JPG e possuem a mesma resolução.

Na avaliação do modelo criado, o conjunto de dados foi separado em grupos de treinamento, testes e validação. Foi utilizado o método de *cross validation*, mais especificamente o *k-fold* para avaliação do modelo. Como citado em Yadav e Shukla (2016), com diferentes distribuições de dados, é possível inferir a qualidade do modelo, podendo mensurar uma média geral de desempenho. Desta forma, o conjunto de dados foi separado em cinco amostras de forma aleatória, garantindo que os conjuntos de treino e de teste possuíssem a mesma distribuição dos dados de cada classe selecionada para o estudo.

2.2. Criação da arquitetura e treinamento do modelo

Após a criação e o tratamento da base de imagens, iniciou-se à fase de criação da estrutura de um modelo visando à classificação dos produtos, sendo utilizado Redes Neurais Convolucionais (CNN). Conforme LeCun et al. (2015), as CNNs são feitas para processar dados que possuem o formato de múltiplos *arrays*, como uma imagem colorida.

Para a elaboração do modelo, foram filmados 100 objetos incluindo tanto objetos semelhantes quanto completamente diferentes entre si, cada um representando um produto específico de uma marca específica. Esses 100 objetos correspondem às 100 classes de produtos definidas no estudo. Para a criação do modelo, a linguagem Python foi utilizada junto da API de redes neurais Keras aliado ao TensorFlow. Por meio de testes para encontrar a melhor arquitetura para o problema, encontrou-se a configuração a seguir.

Modelo de três camadas de convolução e Max *Pooling*, seguido de uma camada de concatenação (*flatten*). Ao final, duas camadas densas foram adicionadas. As camadas de convolução foram configuradas para possuírem 64 filtros. A primeira e a terceira camada possuem kernel de tamanho 3, enquanto a segunda camada possui kernel de tamanho 2. Além disso, a terceira camada foi configurada com a função de ativação *ReLU*. Nas camadas densas, a primeira camada foi configurada com a função de ativação *ReLU* e a segunda camada possui a função de ativação *Softmax* para realizar a classificação final. Com a separação do conjunto de dados e a definição da arquitetura, o modelo pôde ser treinado e avaliado. Esse modelo foi escolhido após diversos testes de arquiteturas serem realizados, tendo se mostrado o mais eficiente para os dados utilizados.

3. Resultados

3.1. Treinamento e avaliação do modelo

O modelo foi treinado com cinco conjuntos de dados diferentes, de acordo com a separação dos dados detalhada na Seção 2.2. Logo, cinco resultados de desempenho foram obtidos. Cada treinamento durou cerca de 10 horas de processamento em um computador com processador Core i5 e com 8 GB de RAM. A Figura 2 mostra o desempenho da acurácia em cada conjunto de dados.

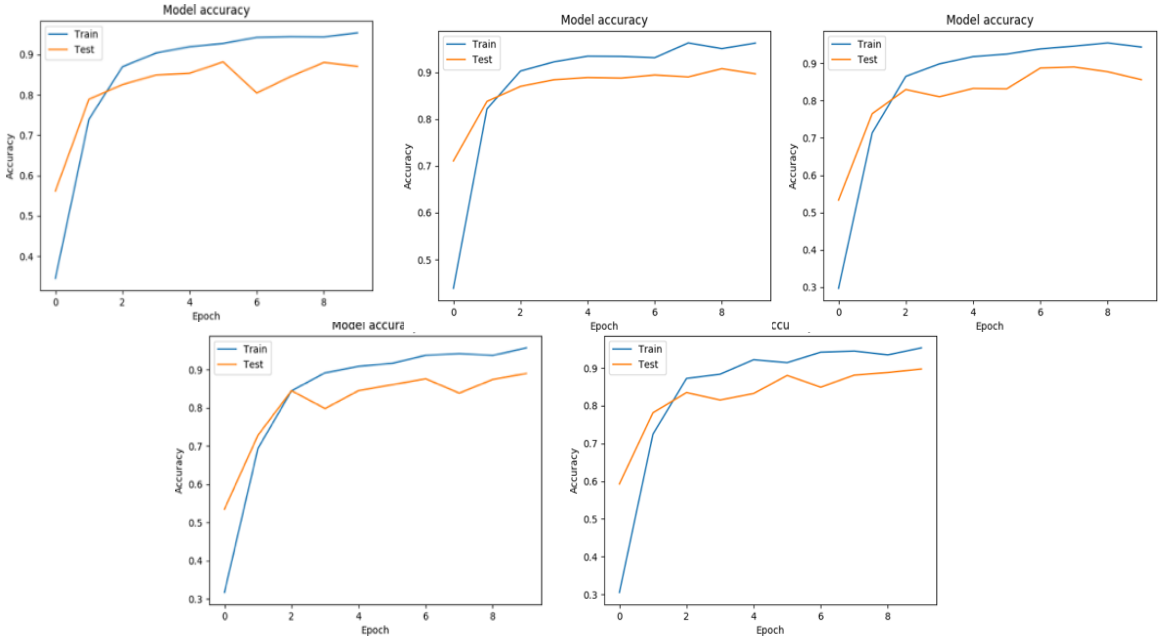

Figura 2. Desempenho durante o treinamento mensurado com a acurácia (*accuracy*) dos cinco conjuntos de dados selecionados.

As acurácias obtidas foram, respectivamente: 88,91%, 89,77%, 87,05%, 85,65% e 89,66%. Em média, o modelo obteve uma acurácia de 88,21%, tendo como melhor resultado o *fold* 2, com 89,77%. Para efeitos de comparação, em Tonioni et al. (2018), o modelo criado tinha o objetivo de classificar produtos em prateleiras de uma mercearia, alcançando uma precisão média de 76,93%. Uma precisão de 45,8% foi alcançada para uma tarefa semelhante em Franco et al. (2017).

3.2. Avaliação do desempenho por classes

Para uma melhor avaliação do banco de imagens criado e do desempenho do modelo, foi realizada uma análise da acurácia da classificação por classe. A Tabela 1 apresenta os melhores desempenhos por classe do modelo treinado, além da quantidade de imagens separadas em cada conjunto de treinamento para cada classe.

Tabela 1. Top-10 classes com melhores desempenhos baseado na acurácia da Classificação

Posição	Classe	Imagens	Acurácia
1	Focinheira Plástica	138	99,71%
2	Shampoo Savana Neutro	176	98,41%
3	Tapete Higiênico c/ 30	111	98,38%
4	Osso Coxa	155	98,32%
5	Rasqueadeira White	110	98,00%
6	Medicamento Biofarm	321	97,94%
7	Brinquedo Pelúcia Girafa	621	97,62%
8	Porta Lixo Osso C/ Refil	140	97,29%
9	Serragem	121	97,19%
10	Brinquedo Durabone	79	96,96%

As classes destacadas com os melhores desempenhos obtiveram uma acurácia superior a 96%. A classe com melhor desempenho foi a Focinheira Plástica (Figura 3), na qual o modelo apresentou erro em apenas 1 imagem, conforme pode ser verificada na Figura 3. O produto possui pouca variação de formato e cor, além de ter poucos produtos semelhantes a ele, o que facilita a identificação.

Figura 3. Exemplo de imagem da classe Focinheira Plástica.

A Tabela 2 apresenta as classes com os piores desempenhos, os quais obtiveram acurácia inferior a 70%. A classe que apresentou o pior desempenho foi a Caixa Transporte Falcon, errando a classificação aproximadamente 60 vezes por treinamento, com uma taxa de sucesso de apenas 19,47%. Foram analisados dois motivos para o desempenho ruim. Conforme a Figura 4, a Caixa de Transporte Falcon possui baixa quantidade de imagens e grande semelhança com a Caixa Transporte Preta (também apresentada na Figura 4).

Tabela 2. Top-10 classes com piores desempenhos baseado na acurácia da classificação.

Posição	Classe	Imagens	Acurácia
1	Caixa Transporte Falcon	75	19,47%
2	Brinquedo Pelúcia Guepardo	97	43,92%
3	Brinquedo Pelúcia Lion	99	44,24%
4	Brinquedo Pelúcia Polvo	94	57,45%
5	Creme Dental Tutti-Frutti	80	60,00%
6	Creme Dental Morango	73	64,38%
7	Brinquedo Pelúcia Dinossauro	121	65,95%
8	Osso Palito Miúdo do Bovino	68	68,24%
9	Osso Palito Frango	74	68,92%
10	Brinquedo Pelúcia Doguinho	111	69,91%

Figura 4. Exemplo de imagem das classes Caixa Transporte Falcon e Caixa Transporte Preta, respectivamente.

4. Considerações Finais.

Este trabalho realiza a tarefa de classificação de produtos de pet shop por meio de técnicas de *deep learning*. As principais contribuições do trabalho foi a criação de um modelo que fosse capaz de classificar corretamente produtos de petshop com acurácia satisfatória, além da criação de uma base de imagens que poderá ser disponibilizada para que outros pesquisadores possam utilizar.

Conforme Melek (2024) existe inúmeros desafios no reconhecimento de produtos, tais como ângulo, iluminação e similaridades das imagens. Entretanto, o modelo criado apresentou um bom desempenho, com acurácia média de 88,21%. Os resultados indicam que o desempenho da classificação deste trabalho não foi influenciado diretamente pela quantidade de imagens por classe, mas sim pela complexidade inerente e pela similaridade entre as classes do problema. Essa conclusão baseia-se na observação de que as taxas de erro foram consistentemente maiores para classes com sobreposição semântica (por exemplo, produtos com embalagens ou cores muito similares), do que para classes com menores números de imagens. Entretanto, mais testes devem ser feitos para considerar fatores diversos que influenciam a acurácia do modelo.

Para trabalhos futuros, possíveis alterações no modelo podem ser realizadas para buscar uma melhora na classificação. De acordo Guimarães et al. (2023), para o reconhecimento de produtos, as características visuais já não são a única opção, podendo aproveitar também de informações textuais para melhoria da acurácia do modelo. Logo, essa abordagem também pode ser utilizada dependendo do propósito da aplicação. Testes utilizando modelos já existentes, como ResNet, deve ser realizado para comparação de desempenho. Outra possibilidade é a utilização de técnicas como *data augmentation* para ampliação da base de imagens. Por fim, fatores como tempo de inferência (classificação em tempo real), eficiência computacional em dispositivos móveis e o uso de GPUs na etapa de treinamento devem ser abordados em trabalhos futuros.

Referências bibliográficas

Franco, A., Maltoni, D., and Papi, S. (2017). Grocery product detection and recognition. *Expert Systems with Applications*, 81:163 – 176.

Guimarães, V., Nascimento, J., Viana, P., and Carvalho, P. (2023). A review of recent advances and challenges in grocery label detection and recognition. *Applied Sciences*,13(5).

Koubaroulis, D., Mata, J., and Kittl, J. (2002). Evaluating Colour-Based Object Recognition Algorithms Using the SOIL-47 Database. In 5th Asian Conference on Computer Vision, Australia.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Melek, C. G., Battini Sonmez, E., and Varlı, S. (2024). Datasets and methods of product recognition on grocery shelf images using computer vision and machine learning approaches: An exhaustive literature review. *Engineering Applications of Artificial Intelligence*, 133:108452.

Merler, M., Galleguillos, C., and Belongie, S. (2007). Recognizing groceries in situ using in vitro training data. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Pietrini, R., Paolanti, M., Mancini, A., Frontoni, E., and Zingaretti, P. (2024). Shelf management: A deep learning-based system for shelf visual monitoring. *Expert Systems with Applications*, 255:124635.

Selvam, P., Faheem, M., Dakshinamurthi, V., Nevgi, A., Bhuvaneswari, R., Deepak, K., and Abraham Sundar, J. (2024). Batch normalization free rigorous feature flow neural network for grocery product recognition. *IEEE Access*, 12:68364–68381.

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*.

Tonioni, A., Serra, E., and di Stefano, L. (2018). A deep learning pipeline for product recognition on store shelves. 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), pages 25–31.

Yadav, S. and Shukla, S. (2016). Analysis of k-fold cross- validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83.