

Comparação de técnicas de transferência de conhecimento em redes convolucionais

De Lima, Matheus

ARTICLE INFO

Keywords:
transfer learning
CNN
fine-tuning
image classification

ABSTRACT

Este artigo visa explorar e comparar técnicas de transferência de conhecimento como transfer learning e fine-tuning aplicadas como métodos para superar obstáculos no desenvolvimento de CNN's para classificação de imagens. O artigo explora um caso de uso e compara as técnicas em domínios próximos e no domínio genérico.

1. Introduction

A falta de dados é um dos maiores desafios na implementação de sistemas de reconhecimento de imagem em diversas áreas, em especial na área médica, a aquisição e classificação de imagens para diagnóstico pode ser extremamente custosa ou até mesmo impossível. Uma alternativa para este problema é a implementação de técnicas de transferência de conhecimento que visam aproveitar o aprendizado adquirido por outras redes para novas tarefas. Essas técnicas são utilizadas e tem se mostrado uma boa opção para remediar a falta de dados, artigos e trabalhos recentes demonstram que se pode obter bons resultados com tais aplicações. No caso de transferências realizadas entre tarefas semelhantes, é demonstrado que o ganho pode ser maior do que a transferência entre aplicações não relacionadas. Este artigo explora e compara a aplicação de transfer learning e fine-tuning para solucionar um problema de classificação de doenças de pele a partir de uma rede treinada na classificação de lesões relacionadas a câncer de pele e uma rede treinada para classificar imagens cotidianas(imagenet).

2. Objectives

Técnicas como transfer learning e fine-tuning já vêm sendo amplamente utilizadas em aplicações que apresentam dificuldade na aquisição de grandes volumes de dados. Um exemplo é a área médica, onde a depender do problema os dados podem ser escassos e expandir os dataset's pode ser inviável. Porém, nem sempre é possível ter um modelo previamente treinado em uma tarefa similar, nesses casos sendo necessário utilizar CNN's treinadas com conjuntos de imagens não relacionadas à área de aplicação em que se deseja trabalhar. O objetivo principal deste trabalho é explorar e comparar o desempenho das técnicas de transfer learning e fine-tuning em domínios semelhantes e diferentes, ou seja, comparar os resultados ao se transferir conhecimento utilizando como base modelos treinados em tarefas similares e modelos treinados em tarefas genéricas, a fim de avaliar as vantagens e desvantagens de tais abordagens.

3. Methodology

3.1. Convolutional Neural Networks

Atualmente, o uso de redes convolucionais é considerado uma das melhores opções para classificação de imagens[5-6] e já é empregado amplamente em imagens médicas. Grande parte disso se dá ao fato de que utilizando camadas convolucionais essas redes são capazes de manter as relações espaciais das imagens durante o processamento[8].

3.2. Transfer Learning

Transfer learning é uma técnica em Deep Learning que consiste em aproveitar o conhecimento adquirido por um modelo previamente treinado para inicializar um segundo modelo em uma tarefa semelhante. Este

ORCID(s):

tipo de abordagem é útil principalmente em casos onde existem poucos dados disponíveis para treinamento[9]. Geralmente, em transfer learning, os pesos das camadas de extração são transferidos para um novo modelo, e um classificador é treinado com um dataset específico, assim é possível aproveitar o conhecimento prévio e ajustar para a tarefa desejada.

3.3. Fine-Tuning

Trata-se de uma técnica de transfer learning onde o conhecimento adquirido é transferido para um novo modelo, a principal diferença na abordagem é que no fine-tuning além do treinamento no classificador, as camadas mais baixas também são treinadas, assim oferecendo uma maior capacidade de ajuste, mas também requerendo mais dados para treinamento[10].

3.4. ResNet50

Trata-se de um rede neural convolucional introduzida pela primeira vez em 2015. Conta com 50 camadas e pertence a família das resnets. Foi projetada para lidar com o problema do gradiente que desaparece em redes profundas. Para isso utiliza blocos residuais que incluem conexões de atalho permitindo que os gradientes fluam diretamente entre camadas distantes[7].

3.5. GradCam

Grad-CAM (Gradient-weighted Class Activation Mapping) consistem em uma técnica de interpretação de redes convolucionais. Funciona gerando uma mapa de calor que usa como base os valores do gradiente da ultima camada de ativação de uma rede em relação a classe apontada, desta forma é possível ver quais regiões da imagem a rede utiliza para tomar as decisões[12].

4. Dataset's

Para a elaboração deste trabalho foram utilizados dois dataset's. O primeiro deles é o "20 Skin Diseases Dataset" que contem cerca de três mil imagens classificadas entre vinte doenças de pele. O dataset apresenta um desbalanceamento considerável, tendo uma classe majoritária que contem mais que o dobro de imagens que as demais. Uma das técnicas utilizadas para lidar com o desbalanceamento foi unificar algumas das classes minoritárias em uma classe chamada "outros", as classes unificadas foram "Urticaria Hives", Vasculitis Photos", "Exanthems and Drug Eruptions", "Melanoma Skin Cancer Nevi and Moles", "Eczema Photos", "Herpes HPV and other STDs Photos". Essas classes apresentavam um número baixo de imagens assim dificultando a aplicação de modelos de classificação. Além disso, foi aplicado um downsampling nas classes majoritárias e um upsampling com datagen nas classes minoritárias a fim de equilibrar o dataset. A geração de dados se baseia em utilizar alterações de zoom, rotação, brilho e crop nas imagens já existentes, essas alterações foram aplicadas separadamente nos conjuntos de treino e teste a fim de evitar vazamento de dados. A finalidade deste dataset é ser utilizado como desafio de solução, ou seja, serão aplicadas técnicas de transferência de conhecimento para desenvolver modelos que visam solucionar este conjunto de dados.

Um problema identificado neste conjunto de dados foi a variedade de composição das fotos, o conjunto possui fotos de várias partes do corpo humano sem classificação de qual parte pertencem e algumas sem um foco no ferimento a ser analisado. Essa variedade pode dificultar a aplicação de tarefas de classificação nas imagens e dificultar a generalização dos modelos, bem como fazer com que estes utilizem elementos incorretos para a extração de informação. Trata-se de um bom dataset para simular o comportamento de usuários de um aplicativo, por exemplo.

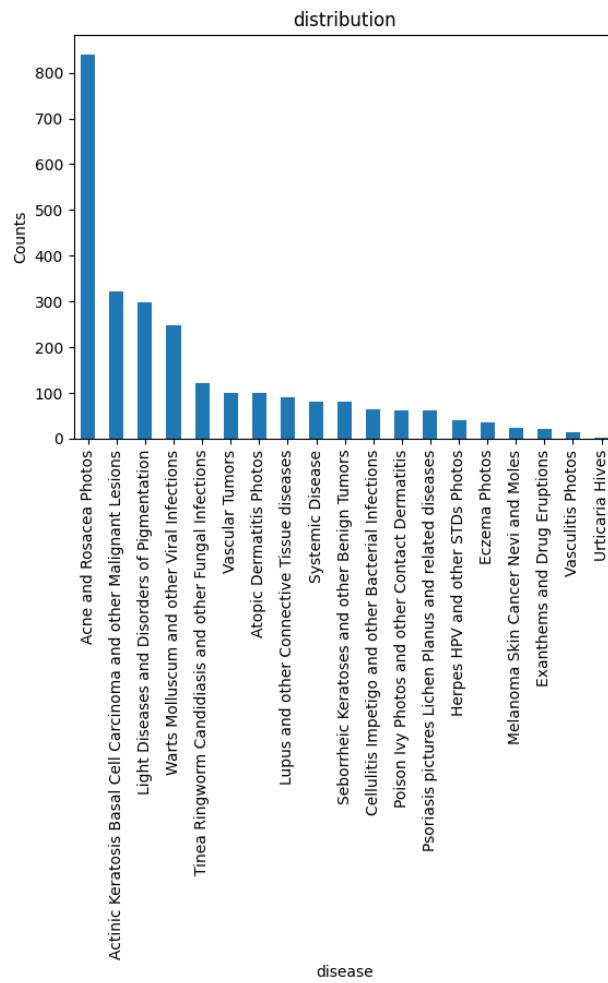


Figure 1: Distribuição de classes do primeiro dataset

Comparação de técnicas de transferência de conhecimento



Figure 2: exemplos de imagens do primeiro dataset

Comparação de técnicas de transferência de conhecimento

O segundo dataset utilizado chama-se "Skin Cancer MNIST: HAM10000" e possui cerca de dez mil imagens classificadas em 7 tipos de câncer de pele. O objetivo deste dataset é ser usado para treinar um modelo que posteriormente terá seu conhecimento transferido para tentar solucionar o dataset anterior. As classes presentes apresentam desbalanceamento e assim como no dataset anterior será aplicada uma rotina de balanceamento que consiste em upsampling como datagen e downsampling a fim de equiparar o número de imagens por classe. Além disso, as classes minoritárias "Vascular lesions", "Dermatofibroma", "Actinic keratoses" foram unificadas, pois apresentavam um número muito baixo de imagens. Este conjunto de dados possui informações sobre a parte do corpo que a lesão estava localizada e estas informações foram utilizadas para remover imagens presentes nas seguintes regiões: 'chest', 'foot', 'unknown', 'neck', 'scalp', 'hand', 'ear', 'genital', 'acral'. As imagens deste dataset estão focadas na lesão e com boa resolução, assim facilitando tarefas de classificação.

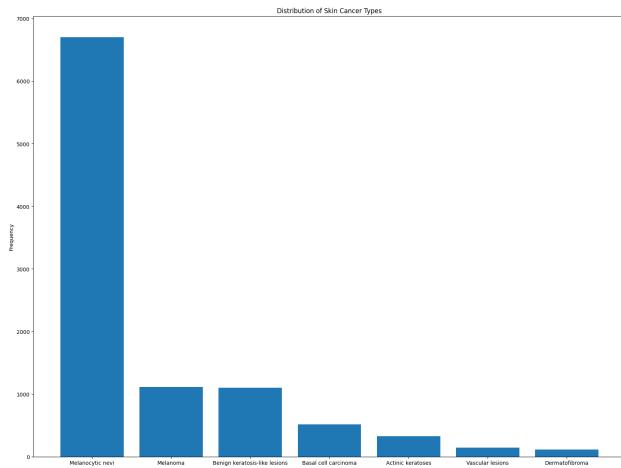


Figure 3: Distribuição de classes do segundo dataset

O terceiro dataset usado foi o "imagenet" que consiste em 14,197,122 imagens. Este dataset não foi utilizado diretamente, porém modelos treinados com ele tiveram seu conhecimento transferido a fim de solucionar o primeiro dataset deste projeto.

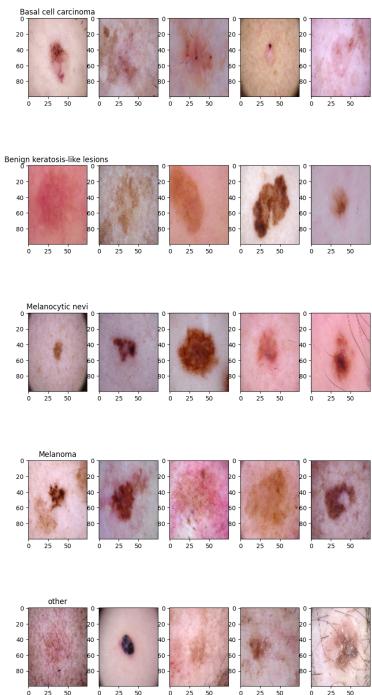


Figure 4: exemplos de imagens do segundo dataset

5. Structure of experiments

Para a realização dos experimentos foram treinados seis modelos, dos quais cinco terão seus resultados comparados.

Inicialmente foi treinado uma rede com o dataset Skin Cancer MNIST: HAM10000, este modelo pretende servir de base tendo seu conhecimento transferido para outros dois.

Após isso foi treinada uma rede no dataset 20 Skin Diseases do zero sem utilizar nenhuma técnica de transferência de conhecimento, o objetivo deste modelo é servir como métrica base para as comparações futuras. Por fim foram treinados quatro modelos utilizando o conhecimento adquirido previamente. Dois destes utilizaram como base o modelo que foi inicialmente treinado para classificar lesões de câncer, tendo um deles utilizado transfer learning e o outro fine-tuning. Dois outros tiveram como base um modelo treinado com imagens diversas do dataset imagenet, divididos da mesma forma que o par anterior. Sendo assim a estrutura dos testes ficou sendo a seguinte:

- **Cenário 1:** Modelo treinado do zero
- **Cenário 2:** Modelo treinado utilizando transfer learning do modelo de câncer
- **Cenário 3:** Modelo treinado utilizando fine-tuning do modelo de câncer
- **Cenário 4:** Modelo treinado utilizando transfer learning do modelo da imagenet
- **Cenário 5:** Modelo treinado utilizando fine-tuning do modelo da imagenet

Com estes cenários é esperado chegar a uma comparação profunda das diferenças de abordagem e suas vantagens e desvantagens.

5.1. Models

Todas as redes treinadas utilizaram a arquitetura resnet50 e tiveram como número limite de épocas 200, com exceção daquela que foi treinada para classificar lesões de câncer que teve o dobro de épocas uma vez que seria usado como base para outros modelos.

Foram utilizados dois callbacks, o EarlyStopping que visa parar o treinamento prematuramente uma vez que o valor da loss pare de diminuir por um determinado período e o ReduceLROnPlateau que reduz a taxa de aprendizado utilizada uma vez que a loss para de evoluir. Para todos os modelos a taxa inicial de aprendizado foi de 0.001 e o otimizador escolhido foi o adamW. Este otimizador implementa melhorias sobre o adam relacionadas ao decaimento de pesos, é indicado para problemas de redes profundas e tem se mostrado eficaz em evitar overfitting[11].

Para treinamento foi utilizado data augmentation utilizando o ImageDataGenerator do keras com os seguintes parâmetros:

- rotation_range=45,
- width_shift_range=0.2,
- height_shift_range=0.2,
- shear_range=0.15,
- zoom_range=0.2,
- horizontal_flip=True,
- vertical_flip=True,
- brightness_range=[0.8,1.2]

A classe `CustomResNet` foi criada e usada como base para todos os modelos usados neste trabalho. Esta classe define uma rede neural baseada na arquitetura **ResNet50**. Pode ser configurada para usar pesos pré-treinados do **ImageNet** ou ser treinada do zero. A rede processa imagens de entrada com tamanho $100 \times 75 \times 3$, utilizando a **ResNet50** sem a camada de classificação (`include_top=False`), Uma vez que usa um classificador próprio.

Após a ResNet50, a rede passa por várias camadas de processamento:

- **Flatten**: Achata a saída da ResNet em um vetor unidimensional.
- **Dropout**: para reduzir overfitting
- **BatchNormalization**: normaliza as ativações
- **Dense**: camada densas com ativação ReLU regularização L2 e 512 neuronios.
- **Dropout**: para reduzir overfitting
- **BatchNormalization**: normaliza as ativações
- **Dense**: camada densas com ativação ReLU regularização L2 e 256 neuronios.
- **Dense (final)**: A camada de saída com ativação softmax, que gera as previsões para a classificação final, com o número de classes especificado (`num_classes`).

Esta rede tem uma arquitetura suficientemente poderosa para solucionar problemas de classificação de imagens e pode facilmente ser customizada.

6. Results

Essa seção irá apresentar os resultados encontrados nos experimentos.

6.1. Rede treinada no dataset Skin Cancer MNIST: HAM10000

A primeira rede treinada com o dataset de imagens de câncer treinou por 98 épocas sendo interrompida de forma prematura pelo callback, pois não estava mais evoluindo. Esta rede alcançou os seguintes valores:

| Métrica | Treinamento | Validação | Teste |
|-----------------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.9737 | 0.8400 | 0.6200 |
| F1 Score (média) | 0.9735 | 0.8424 | 0.6183 |
| Loss | 0.1275 | 0.7299 | 1.6676 |
| Recall | 0.9707 | 0.8380 | 0.6080 |

Table 1: Resultados de Avaliação do Modelo

Apesar do uso de técnicas para remediar o overfitting como uso de regularização, dropout e data augmentation a rede ainda assim apresentou problemas neste quesito finalizando com um desempenho de teste inferior ao apresentado durante o treinamento.

É possível analisando a figura 6 (matriz de confusão da rede treinada no primeiro dataset) notar a diagonal principal bem destacada como o esperado, porém o modelo apresentou muitos focos de confusão sendo o principal entre a classe Basal cell carcinoma e melanocyte nevi o que pode indicar um alto indicie de similaridade entre as imagens no dataset.

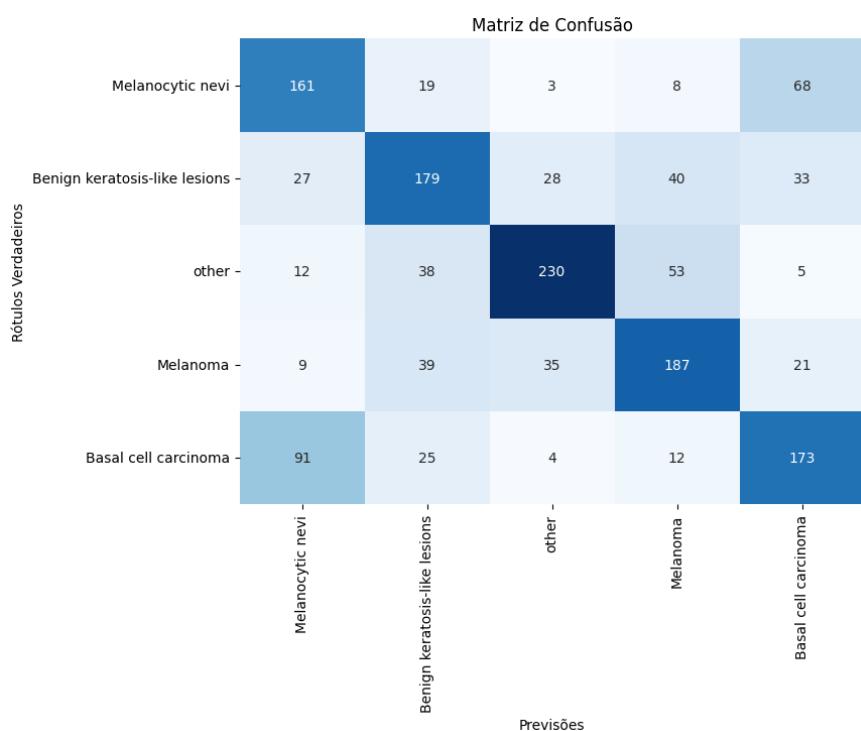


Figure 5: matriz de confusão da rede treinada no primeiro dataset

6.2. Cenário 1

Este modelo foi treinado do zero no dataset de doenças de pele, após a sanitização de dados este dataset ficou com 14 classes. A rede treinou por 90 épocas quando foi interrompida pelo callback. O seu desempenho foi o seguinte: Esse modelo apresentou uma baixa precisão e um alto índice de overfitting, sua matriz de

| Métrica | Treinamento | Validação | Teste |
|-----------------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.9748 | 0.7449 | 0.3143 |
| F1 Score (média) | 0.9746 | 0.7534 | 0.3133 |
| Loss | 0.3148 | 1.3536 | 4.4203 |
| Recall | 0.9627 | 0.7245 | 0.2905 |

Table 2: Resultados de Avaliação do Modelo

confusão foi a seguinte:

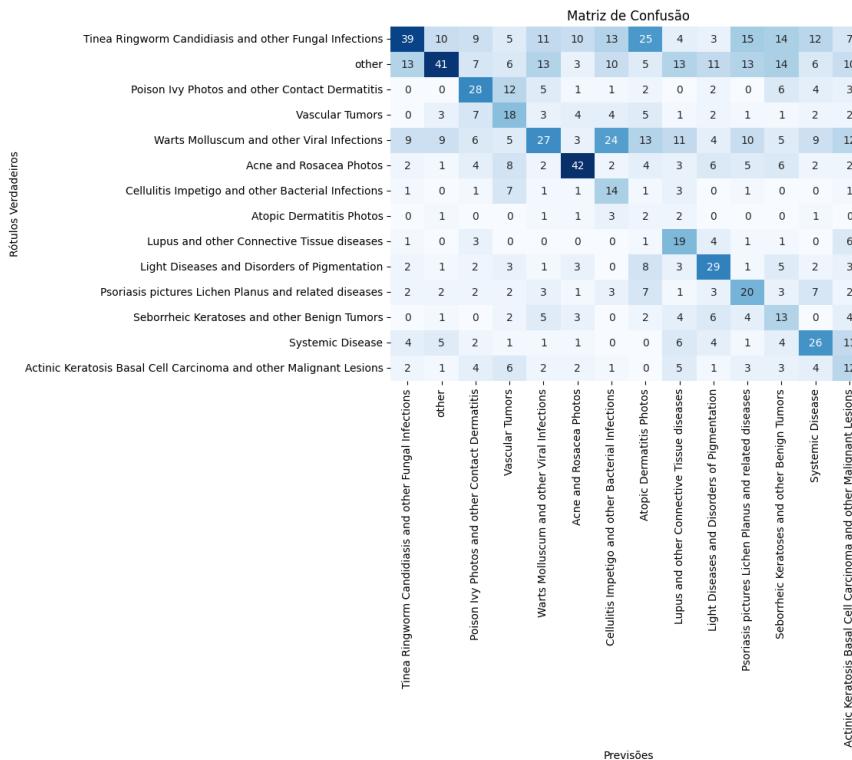


Figure 6: matriz de confusão do Cenário 1

Comparação de técnicas de transferência de conhecimento

O desempenho da rede foi baixo tendo sido principalmente influenciado pelas classes: "Cellulitis Impetigo and other Bacterial Infections", "Actinic Keratosis Basal Cell Carcinoma and other Malignant Lesions", "Seborrheic Keratoses and other Benign Tumors" e "Atopic Dermatitis Photos" tendo essa ultima atingido apenas 2% de precisão, desconsiderando essas classes a precisão do modelo seria de 40%. As classes "Acne and Rosacea Photos", "Other" e "Tinea Ringworm Candidiasis and other Fungal Infections" apresentaram um desempenho superior a 50%. Algumas classes foram confundidas sendo elas "Vascular Tumors" com "Light Diseases and Disorders of Pigmentation" e "Actinic Keratosis Basal Cell Carcinoma and other Malignant Lesions" confundida com "Seborrheic Keratoses and other Benign Tumors".

6.3. Cenário 2

Esta rede foi treinada usando o dataset "20 skin diseases dataset" após ter tido o conhecimento da rede treinada no dataset "Skin Cancer MNIST: HAM10000" transferido. Esse processo se deu utilizando as funcionalidades "set_weights" e "get_weights" do tensorflow, os pesos das camadas mais baixas (extração de features) foram congelados. O modelo treinou por 90 épocas quando foi interrompido pelo callback e seu desempenho foi o seguinte:

| Métrica | Treinamento | Validação | Teste |
|----------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.5675 | 0.5082 | 0.2228 |
| F1 Score (média) | 0.5603 | 0.5039 | 0.215 |
| Loss | 2.2065 | 2.3922 | 3.5192 |
| Recall | 0.2784 | 0.2469 | 0.3114 |

Table 3: Resultados de Avaliação do Modelo

A rede apresentou overfitting e um desempenho inferior ao da rede treinada do zero. Analisando a matriz de confusão fica evidente que o modelo foi pior do que aquele treinado do zero.



Figure 7: matriz de confusão do Cenário 2

6.4. Cenário 3

Esta rede foi treinada usando o dataset "20 skin diseases dataset" após ter tido o conhecimento da rede treinada no dataset "Skin Cancer MNIST: HAM10000" transferido. O processo ocorreu como no cenário anterior porém as camadas mais baixas não foram congeladas. A rede treinou por 90 épocas até ser interrompida pelo callback e seu desempenho foi o seguinte:

| Métrica | Treinamento | Validação | Teste |
|----------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.9709 | 0.7714 | 0.3333 |
| F1 Score (média) | 0.9704 | 0.7784 | 0.3232 |
| Loss | 0.3128 | 1.2352 | 4.3717 |
| Recall | 0.9585 | 0.7531 | 0.3114 |

Table 4: Resultados de Avaliação do Modelo

A rede apresentou overfitting como as outras e obteve resultados similares aos do cenário 1. A matriz de confusão pode ser vista abaixo:

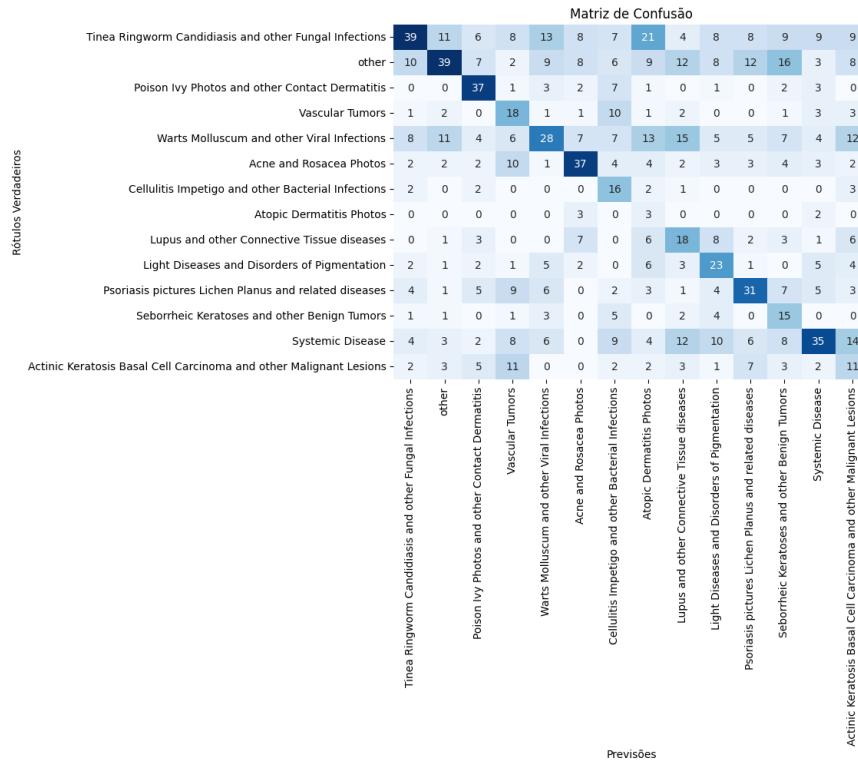


Figure 8: Matriz de confusão do Cenário 3

6.5. Cenário 4

O cenário 4 foi implementado assim como o cenário 2 porém ao invés de utilizar os pesos do modelo treinado com o dataset de imagens de câncer de pele utilizou os pesos da rede treinada com imagenet disponíveis no tensorflow. O desempenho do modelo foi o seguinte:

O desempenho foi similar ao dos outros cenários com a diferença de que os erros foram ligeiramente mais variados.

| Métrica | Treinamento | Validação | Teste |
|-----------------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.8641 | 0.7286 | 0.3428 |
| F1 Score (média) | 0.8628 | 0.7451 | 0.3345 |
| Loss | 1.1590 | 1.5493 | 3.2979 |
| Recall | 0.7455 | 0.6408 | 0.2628 |

Table 5: Resultados de Avaliação do Modelo

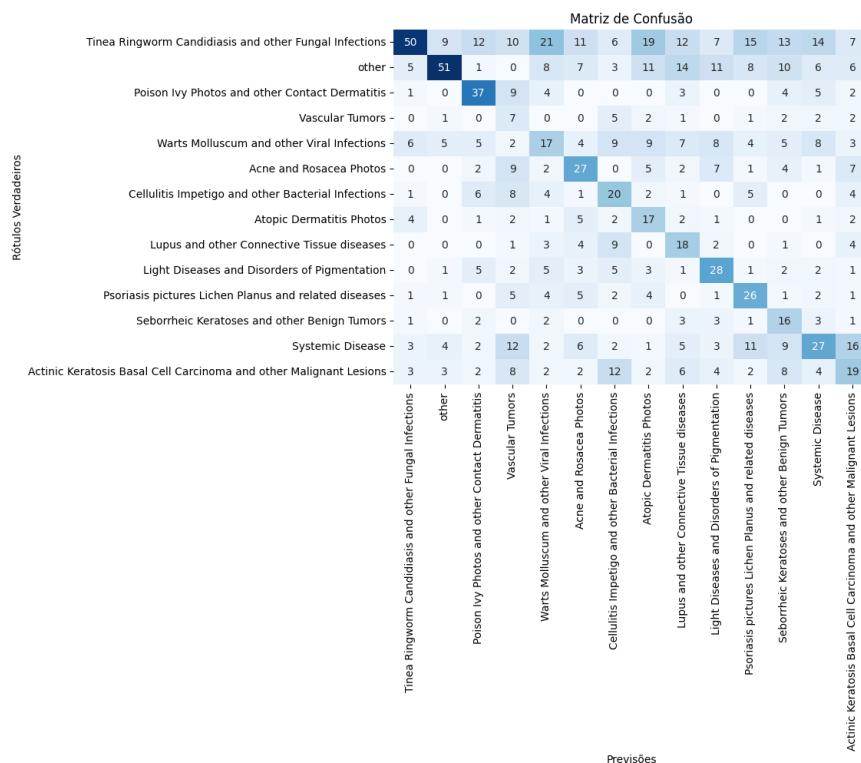


Figure 9: Matriz de confusão do Cenário 4

6.6. Cenário 5

O cenário 5 foi implementado de forma similar ao cenário 3 porém o fine-tuning usou como base o modelo treinado com o imagenet. Os resultados encontrados foram os seguintes:

| Métrica | Treinamento | Validação | Teste |
|-----------------------------|-------------|-----------|--------|
| Categorical Accuracy | 0.9898 | 0.8347 | 0.4409 |
| F1 Score (média) | 0.9897 | 0.8444 | 0.3345 |
| Loss | 0.1764 | 0.8011 | 3.2979 |
| Recall | 0.9866 | 0.8245 | 0.4199 |

Table 6: Resultados de Avaliação do Modelo

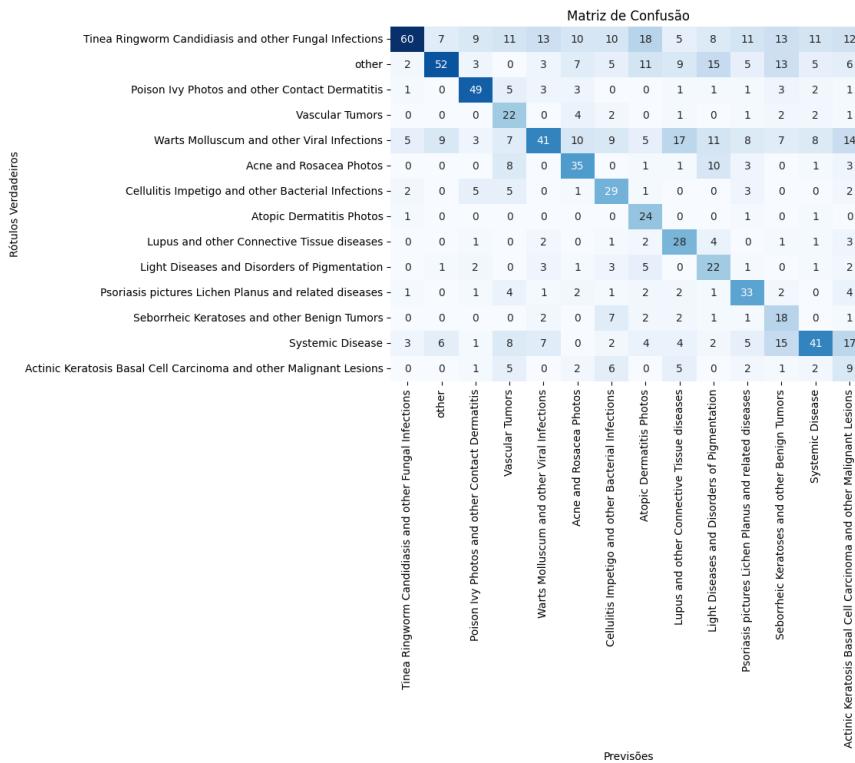


Figure 10: Matriz de confusão do Cenário 4

Apesar de também ter apresentado overfitting este modelo teve o desempenho superior aos demais e ao desconsiderar as 4 classes que menos performaram ele chega na precisão de 54.67%.

A comparação final de abordagem das redes(em teste) ficou da seguinte forma:

| Métrica | cenário 1 | cenário 2 | cenário 3 | cenário 4 | cenário 5 |
|-----------------------------|-----------|-----------|-----------|-----------|-----------|
| Categorical Accuracy | 0.3143 | 0.2228 | 0.3333 | 0.3428 | 0.4409 |
| F1 Score (média) | 0.3133 | 0.215 | 0.3232 | 0.3345 | 0.3345 |
| Loss | 4.4203 | 3.5192 | 4.3717 | 3.2979 | 3.2979 |
| Recall | 0.2905 | 0.3114 | 0.3114 | 0.2628 | 0.4199 |

Table 7: Comparação de cenários

7. Conclusion

O cenário 1 apresenta valores baixos nas métricas de avaliação. Este modelo foi treinado do zero e assim sendo depende exclusivamente dos dados e de sua arquitetura para evoluir, por se tratar de um dataset pequeno e com alta variabilidade de imagens o modelo acaba tendo graves problemas.

O cenário 2 apresenta o pior desempenho entre todos isso pode indicar que o conhecimento adquirido do modelo de câncer não se traduz de forma eficiente para o dataset de doenças de pele. Uma das principais causas deste problema pode ser a diferença no tipo de imagens, o dataset de câncer tem imagens focadas na lesão e padronizadas enquanto o dataset de doenças de pele apresenta uma variedade muito grande de imagens e elementos presentes. Por se tratar de transfer learning as camadas de extração não se adaptam as necessidades da tarefa pretendida.

No terceiro cenário as métricas melhoraram um pouco, isso pode ter ocorrido por se tratar de fine-tuning sendo assim as camadas mais baixas não estão congeladas e podem se adequar as peculiaridades da tarefa alvo, mesmo assim desempenho não teve grande diferença quando comparada ao cenário 1.

No quarto cenário o desempenho aumenta em relação aos anteriores, mas o recall é baixo o que indica que o modelo pode ter tido dificuldades em se ajustar corretamente.

O último cenário teve o melhor desempenho entre todos apresentando valores melhores em todas as métricas utilizadas.

O uso de redes treinadas em domínio geral se mostrou superior ao uso de redes treinadas com imagens específicas, esse fato pode ter ocorrido devido ao dataset da tarefa alvo possuir imagens muito variadas assim favorecendo a transferência de conhecimento que se deu a partir de redes mais generalistas, isso pode ser notados analisando os cenários 2 e 3 contra 4 e 5. O uso de fine-tuning mostrou um desempenho superior ao de transfer learning uma das causas pode ter sido a alta variabilidade das imagens da tarefa alvo que acabam por exigir ajustes na camada de extração para atingirem melhores resultados.

8. GradCam application

Ao final foi utilizada a técnica de gradcam para visualizar as regiões de interesse dos diferentes modelos treinados. com essa abordagem é possível identificar quais itens das imagens cada modelo está considerando para suas classificações. O resultado dessa aplicação pode ser visto abaixo:



Figure 11: Imagens originalmente

Os resultados foram bastante diversos em todos os modelos, tendo imagens em que a ativação foi condizente com o esperado e imagens onde a região de interesse foi completamente ignorada, isso se deve a grande diversidade do dataset e baixo número de imagens o que dificulta a formação de padrões eficientes.

Comparação de técnicas de transferência de conhecimento

do zero

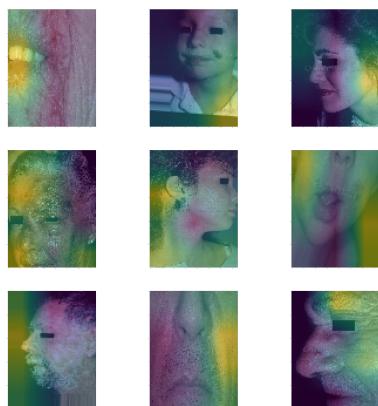


Figure 12: Visão do modelo do cenário 1

transfer cancer



Figure 13: Visão do modelo do cenário 2

fine tuning cancer

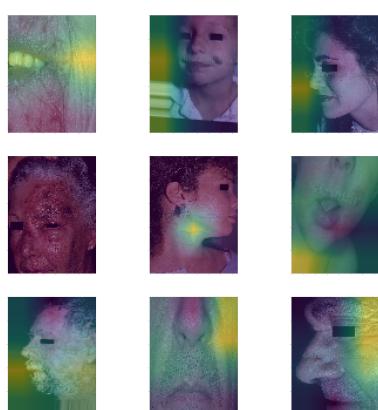
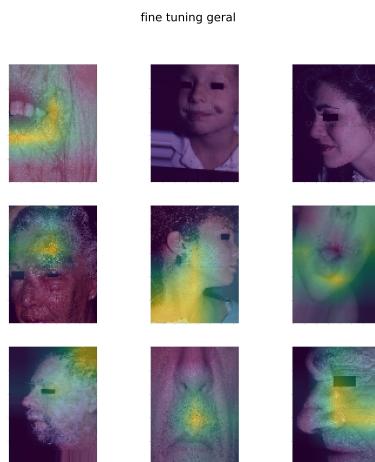


Figure 14: Visão do modelo do cenário 3

**Figure 15:** Visão do modelo do cenário 4**Figure 16:** Visão do modelo do cenário 5

9. Future Directions

Neste trabalho, foram exploradas diferentes abordagens de implementação de CNN's para lidar com dataset's pequenos tais como transfer learning e fine-tuning e áreas específicas e no domínio geral. Alguns dos possíveis próximos passos a serem tomados são: **Interpretabilidade:** Expandir o uso de gradcam e explorar de forma mais extensiva os dataset usado visando dar um melhor direcionamento as estratégias de treino. **Expansão dos dataset's:** Buscar formas de expandir os dataset's utilizados ou reproduzir os experimentos com dataset's maiores afim de comparar o desempenho e os impactos no aumento do numero de imagens disponíveis. **Técnicas de otimização:** Investigar e aplicar técnicas mais avançadas de otimização como **técnicas de aprendizado diferenciáveis**[13] e **pipelines de data augmentation**[14].

Comparação de técnicas de transferência de conhecimento

1. K. Mader. Skin Cancer MNIST: HAM10000. Kaggle. Disponível em: <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/code>. Acesso em: 9 dez. 2024.
2. ALAM, Haroon. 20 Skin Diseases Dataset. Kaggle. Disponível em: <https://www.kaggle.com/datasets/haroonalam16/20-skin-diseases-dataset>. Acesso em: 9 dez. 2024.
3. DENG, Jia et al. ImageNet: A Large-Scale Hierarchical Image Database. 2009. Disponível em: <https://www.image-net.org>. Acesso em: 9 dez. 2024.
4. Alzubaidi, L.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Duan, Y. Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model. *Electronics* 2020, 9, 445.
5. Alzubaidi, L.; Fadhel, M.A.; Olewi, S.R.; Al-Shamma, O.; Zhang, J. DFUQUTNet: Diabetic foot ulcer classification using novel deep convolutional neural network. *Multimed. Tools Appl.* 2020, 79, 15655–15677.
6. DENG, Jia et al. ImageNet: A Large-Scale Hierarchical Image Database. 2009. Disponível em: <https://www.image-net.org>. Acesso em: 9 dez. 2024.
7. HE, Kaiming et al. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 770-778. Disponível em: <https://arxiv.org/abs/1512.03385>. Acesso em: 9 dez. 2024.
8. LeCun, Yann et al. Gradient-Based Learning Applied to Document Recognition. In: Proceedings of the IEEE, 1998, vol. 86, no. 11, p. 2278-2324. DOI: 10.1109/5.726791.
9. Pan, Sinno Jialin; Yang, Qiang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, vol. 22, no. 10, p. 1345-1359. DOI: 10.1109/TKDE.2009.191.
10. Yosinski, Jason et al. How Transferable Are Features in Deep Neural Networks? In: Advances in Neural Information Processing Systems (NeurIPS), 2014, p. 3320-3328. Disponível em: <https://arxiv.org/abs/1411.1792>.
11. Loshchilov, Ilya; Hutter, Frank. Decoupled Weight Decay Regularization. In: Proceedings of the International Conference on Learning Representations (ICLR), 2019. Disponível em: <https://arxiv.org/abs/1711.05101>.
12. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). Disponível em: <https://arxiv.org/abs/1610.02391>.
13. Differential Learning Rates Smith, L. N. (2017). Cyclical learning rates for training neural networks. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 464-472. Disponível em <https://doi.org/10.1109/WACV.2017.51>
14. Data Augmentation Pipelines Perez, L., Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit.* Disponível em <https://arxiv.org/abs/1712.04621>