Detecção de pneumonia usando redes neurais convolucionais treinadas com destilação do conhecimento obscuro

José Vitor Santos Silva, Leonardo Nogueira Matos

Departamento de Computação – Universidade Federal de Sergipe (UFS)
São Cristóvão – SE – Brasil

{jose.silva,leonardo}@dcomp.ufs.br

Abstract. *Pneumonia is defined as inflammation of the lung tissue due to an infectious agent, which can be caused by a virus or bacteria. It is a common infection and affects mainly the elderly and children. A chest X-ray is often the best way to diagnose the disease. In this work, X-ray images are used for diagnosing patients (normal health or pneumonia) by applying convolutional neural networks trained with the dark knowledge distillation technique. This training technique is able to improve the performance of a neural network, allowing a network of simple architecture and less computationally expensive (apprentice model) to achieve a performance close to a more complex model, called the teacher model. The results obtained demonstrate the viability of the technique.*

Resumo. *A pneumonia é definida como inflamação do tecido pulmonar devido a um agente infeccioso, pode ser causada por um vírus ou por uma bactéria. É uma infecção comum e afeta principalmente idosos e crianças. Atualmente, o exame de raio X do tórax é frequentemente a melhor maneira de diagnosticar a doença. Nesse trabalho, imagens de raio X são usadas como base para diagnóstico de pacientes (saúde normal ou com pneumonia) aplicando-se para isso redes neurais convolucionais treinadas com a técnica de destilação do conhecimento obscuro. Esta técnica de treinamento é capaz de melhorar o desempenho de uma rede neural, permitindo que uma rede de arquitetura simples e menos custosa computacionalmente (modelo aprendiz) possa atingir um desempenho próximo ao de um modelo mais complexo, chamado de modelo professor. Os resultados obtidos demonstram a viabilidade da técnica.*

1. Introdução

A pneumonia é definida como inflamação do tecido pulmonar devido a um agente infeccioso, pode ser causada por um vírus ou por uma bactéria. É uma infecção comum e afeta principalmente idosos e crianças. O exame de raio X do tórax é frequentemente a melhor maneira de diagnosticar a doença [WHO 2001], desempenhando um papel importante no atendimento clínico [Franquet 2001]. Neste trabalho, imagens de raio X são usadas como base para diagnóstico de pacientes (saúde normal ou com pneumonia) aplicando-se para isso redes neurais convolucionais treinadas com a técnica de destilação do conhecimento obscuro [Hinton, Vinyals e Dean 2015].

A evolução do aprendizado profundo, do inglês *deep learning*, possibilitou o surgimento de modelos cada vez mais poderosos. Muitos desses modelos obtiveram sucesso em diversas aplicações práticas, inclusive no campo da bioinformática

[Goodfellow et al. 2016, Längkvist et al. 2014]. Além disso, alguns trabalhos já utilizaram modelos de aprendizado profundo para o diagnóstico de pneumonia com base em imagens de raio X [Rajpurkar et al. 2017, Kermany et al. 2018, Chouhan et al. 2020].

Apesar desse sucesso, os modelos também cresceram em complexidade e custo computacional, o que limita a utilização dos mesmos em algumas situações. Por outro lado, os modelos mais simples não conseguem, na maioria das vezes, atingir um desempenho tão bom quanto um modelo de arquitetura profunda. Diante disso, são buscadas alternativas que aproveitem a capacidade dos modelos profundos e que ao mesmo tempo possam mitigar suas limitações. Nesse trabalho, utilizamos a técnica de destilação do conhecimento obscuro com o propósito de treinar uma rede de arquitetura simples que possa atingir um desempenho próximo ao de um modelo mais complexo. A principal ideia dessa técnica é que, após obter um modelo, a partir de um agrupamento de máquinas de aprendizado ou de uma única isolada com arquitetura profunda (modelo professor), um novo treinamento é realizado para transferir o conhecimento para um outro, formado por uma única máquina com arquitetura mais simples (modelo estudante).

A base de dados usada nesse trabalho *Chest X-Ray Images (Pneumonia)* [Kermany et al. 2018], está disponível publicamente e consiste em imagens de raio X do tórax categorizadas em três classes: normal, pneumonia viral e pneumonia bacteriana. Os modelos escolhidos para classificação foram as redes neurais convolucionais, visto que nos últimos anos este tipo de rede neural obteve bastante sucesso em problemas de visão computacional.

2. Revisão da Literatura

2.1. Redes Neurais Convolucionais

Uma rede neural convolucional é uma rede neural artificial que aplica uma operação de convolução aos dados de entrada. Uma camada convolucional aplica uma matriz de pesos, chamada de filtro ou *kernel*, ao longo da imagem (percorre a imagem, aplicando operações de multiplicação e soma, vide Figura 1), esses filtros fazem a extração de características, identificando curvas, linhas, formas e outros padrões. Os filtros normalmente são inicializados de forma aleatória e são aprendidos durante o treinamento pela rede neural. O tamanho do kernel e do passo entre as multiplicações (*stride*) são alguns dos parâmetros definidos pelo projetista deste tipo de rede.

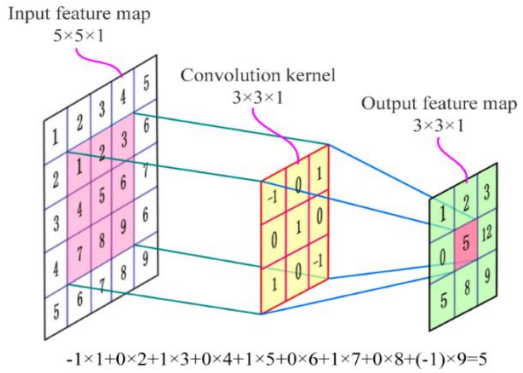

Figura 1. Diagrama esquemático da operação de convolução utilizando um kernel 3x3. Fonte: [Gong et al. 2019].

Conforme a rede fica mais profunda características mais específicas são extraídas, camadas mais rasas detectam formas, curvas, linhas e padrões simples enquanto as mais

profundas detectam características mais complexas como olhos, rostos e outros tipos de padrões.

2.2. Destilação do Conhecimento

A destilação do conhecimento é uma técnica utilizada para melhorar a capacidade de generalização de uma rede neural. O processo visa transferir o conhecimento de um modelo para outro, sendo feito geralmente de um modelo profundo (modelo professor) para um de arquitetura mais simples (modelo estudante).

Existem algumas maneiras de se destilar o conhecimento. Uma abordagem, proposta por [Ba e Caruana 2014] é a utilização das *logits* (entradas da função softmax) de modo que o modelo estudante, ou aprendiz, tente aproximar as suas *logits* das do modelo professor. Outra abordagem é a chamada destilação do conhecimento obscuro, do inglês *dark knowledge distiller*, proposta por Hinton et al. (2015). Nesta outra abordagem o modelo estudante recebe um conjunto de probabilidades geradas pelo modelo professor. Para tal, aplica-se nas *logits* do modelo professor a função softmax, modificada pela introdução de um coeficiente de suavização. Em seguida, durante o treinamento, o modelo estudante tenta aproximar as probabilidades geradas a partir das próprias *logits* das probabilidades geradas pelo modelo professor, chamadas de *soft targets*. Neste artigo utilizaremos a destilação do conhecimento obscuro. Usaremos o termo aprendiz como referência ao modelo com arquitetura mais simples e, para diferenciar a forma como ele é treinado, acrescentaremos quando necessário a designação "treinamento convencional" e "destilação de conhecimento obscuro". O coeficiente T incorporado à função softmax foi chamado por Hinton de temperatura, vide equação 1.

$$Softmax(x_i) = \frac{e^{x_i/T}}{\sum_j e^{x_j/T}} \quad (1)$$

Onde x é o vetor contendo as *logits* para uma determinada entrada e os subscritos i e j são os índices. A função sofmax em sua forma padrão (T=1) gera probabilidades referentes às classes do problema, de modo que uma das classes terá um valor muito próximo de 1 e as restantes valores muito próximos de 0. Isto não é muito diferente das chamadas *hard targets* usadas nos meios convencionais de treinamento. Com o aumento do coeficiente T a distribuição da probabilidade entre as classes fica mais suave. Essa alteração possibilita que o modelo estudante utilize uma distribuição probabilística que conserve não apenas informações de uma das classes, mas também das outras. Nos métodos convencionais de treinamento, essa informação, isto é, os escores das classes perdedoras, não é conhecida, por isso a técnica se chama destilação do conhecimento obscuro.

O uso das *soft targets* cria uma noção de ranqueamento entre as classes. Por exemplo, dada uma amostra de treinamento com *hard target* [1,0,0], não é possível extrair nenhuma informação de ranqueamento entre as classes 2 e 3. Contudo, se utilizarmos *soft targets*, por exemplo [0.8, 0.15, 0.5] a informação de ranqueamento entre as classes começa a aparecer. Se aumentarmos o valor de T isso será ainda mais evidente [Tang, Wang e Zhang 2016]. Além de prover mais informação ao modelo aprendiz, o uso das *soft targets* torna o treinamento mais rápido, uma vez que resulta numa menor variação do gradiente entre as amostras de treinamento, possibilitando que o modelo aprendiz possa ser treinado com menos dados e frequentemente com uma taxa de aprendizado maior [Hinton, Vinyals e Dean 2015].

Durante o treinamento usando destilação do conhecimento obscuro, a função de perda, que indica o quão perto a saída do modelo está do valor correto, é calculada usando a saída do modelo aprendiz, as *soft targets* geradas pelo modelo professor e as *hard targets* (em algumas variações o treinamento é feito usando apenas a saída do modelo aprendiz e as *soft targets*). É calculada assim, uma função de perda que é a soma ponderada de duas componentes. A primeira, vide equação 2, consiste na entropia cruzada das probabilidades do modelo professor e das probabilidades do modelo aprendiz, aplicando-se às *logits* de ambos os modelos a função softmax, com o mesmo valor de T. A segunda componente, vide equação 3, consiste na entropia cruzada das *hard targets* e das probabilidades do modelo aprendiz (aplicando-se a softmax com T=1). A função de perda resultante pode ser vista na equação 4:

$$L_1 = H\big(\sigma(z_t; T = \tau), \sigma(z_s; T = \tau)\big) \quad (2)$$

$$L_2 = H\big(y, \sigma(z_s; T = 1)\big) \quad (3)$$

$$L = \alpha * L_1 + \beta * L_2 \quad (4)$$

Onde z_s e z_t são as *logits* do modelo professor e do modelo aprendiz, respectivamente. y são as *hard targets*. H é a função entropia cruzada. σ é a função softmax parametrizada pela temperatura T. α e β são os pesos. Nota-se, portanto, que essa técnica de treinamento introduz três novos hiper parâmetros: T, α e β. Hinton usou nos experimentos valores de T entre 1 e 20 e obteve resultados melhores com β consideravelmente menor que α. O gradiente, durante o treinamento, é calculado com base nessa função.

3. Metodologia

A metodologia empregada nesse trabalho contempla três fases: a primeira envolve a definição, treinamento e análise do desempenho dos modelos professor e aprendiz. A segunda envolve a comparação dos resultados do modelo aprendiz com os resultados do modelo professor, bem como a análise da eficácia da técnica de destilação do conhecimento obscuro. Por fim, a terceira fase envolve a comparação dos resultados obtidos com outros trabalhos da literatura.

Os testes iniciais dos modelos propostos foram feitos com a base de dados *Chest X-Ray Images (Pneumonia)* [Kermany et al. 2018] que está disponível publicamente. Esta base de dados possui três classes: imagens de raio X do tórax de pessoas saudáveis, de pessoas com pneumonia bacteriana e de pessoas com pneumonia viral, vide Figura 2. Em nossos experimentos agrupamos as duas classes de pneumonia (bacteriana e viral) em uma só, de modo que o problema passou a ser de classificação binária. O conjunto de dados possui ao todo 5840 imagens, sendo 1575 de pessoas saudáveis e 4265 de pessoas com pneumonia viral ou bacteriana. A Tabela 1 apresenta as partições da base de dados usadas nos experimentos para treinamento e teste do modelo, usamos as mesmas partições que os outros trabalhos da literatura usaram, para que a comparação dos resultados fosse justa.

Além disso, foram usadas técnicas de *data augmentation* e regularização L2 ($\lambda = 0.000125$) para mitigar os efeitos do *overfitting*. As transformações de *data augmentation* usadas foram flip horizontal aleatório (para a rede aprender a lidar com sintomas da pneumonia em ambos os pulmões) e redimensionamento com corte aleatório

(para a rede aprender a associar uma gama mais ampla de ativações espaciais a uma determinada classe).

Figura 2. Exemplos de amostras da base de dados.

Tabela 1. Partições do conjunto de dados usadas para treinamento e teste.

	Treinamento (número de imagens)	Teste (número de imagens)
Normal	1341	234
Pneumonia	3875	390
Total	5216	624

A arquitetura escolhida para o modelo professor foi um agrupamento de redes, similar ao que foi feito por Chouhan et al. (2020). Foram usadas cinco arquiteturas de modelos previamente treinados com o conjunto de dados ImageNet [Russakovsky 2015] (uma base de dados com mais de 14 milhões de imagens com 1000 classes), esse treinamento tornou esses modelos capazes de identificar diversos tipos de padrões. Os modelos escolhidos para o agrupamento foram: AlexNet [Krizhevsky, Sutskever e Hinton 2012], DenseNet121 [Huang et al. 2017], ResNet18 [He et al. 2016], VGG19 [Simonyan e Zisserman 2014], e GoogLeNet [Szegedy et al. 2015].

Em alguns problemas de visão computacional uma boa abordagem é usar modelos previamente treinados, a vantagem disso é que podemos tornar o treinamento mais rápido. A técnica de aproveitar os pesos de uma rede neural e ajustá-la para um domínio diferente do qual ela foi originalmente treinada se chama transferir aprendizado, do inglês *transfer learning*. Para o treinamento dos modelos do agrupamento (modelo professor) usamos essa técnica, para tal, a última camada linear de todos os modelos foi modificada para as duas classes da nossa base de dados. Além disso, os pesos das camadas iniciais dos modelos foram 'congelados', isto é, esses pesos não foram atualizados durante o treinamento. Nos modelos AlexNet, DenseNet121, ResNet18 e GoogLeNet as camadas congeladas foram as mesmas das do trabalho de Chouhan et al. (2020). Já para o modelo VGG19, congelamos as camadas convolucionais do extrator de características e atualizamos apenas os pesos do classificador.

Para o modelo aprendiz foi escolhida uma arquitetura simples, com apenas quatro camadas convolucionais, ele também tinha camadas de *maxpooling* e *batch normalization*. Destaca-se a arquitetura muito mais simples do modelo aprendiz em relação ao modelo professor (agrupamento de redes), possuindo uma quantidade de parâmetros treináveis menor e exigindo menos processamento. Em todos os modelos o treinamento foi feito com o otimizador Adam [Kingma e Ba 2014], com uma taxa de aprendizagem de 0.001 que a cada três épocas era dividida por dois. Em todos os casos o tamanho do *batch* usado foi 32. No treinamento convencional a função de perda utilizada foi a entropia cruzada.

A verificação da eficácia da técnica de destilação do conhecimento obscuro foi feita por meio da comparação do desempenho do modelo aprendiz em relação ao do modelo professor, considerando também o desempenho do modelo aprendiz treinado sem a técnica de destilação do conhecimento obscuro. Isto é, comparamos o modelo mais simples (aprendiz) treinado com os dados originais, sem interferência de um modelo professor. Por fim, os resultados obtidos são comparados aos de outros trabalhos que usaram a mesma base de dados.

4. Resultados

4.1. Treinamento e análise do desempenho dos modelos professor e aprendiz

O treinamento de cada um dos modelos do agrupamento (modelo professor) foi feito durante 50 épocas, durante o treinamento desses modelos foi adicionado ruído gaussiano às amostras de treinamento para melhorar a capacidade de generalização e evitar o *overfitting*. A Figura 3 apresenta a evolução da acurácia e *loss* durante o treinamento de cada um dos modelos.

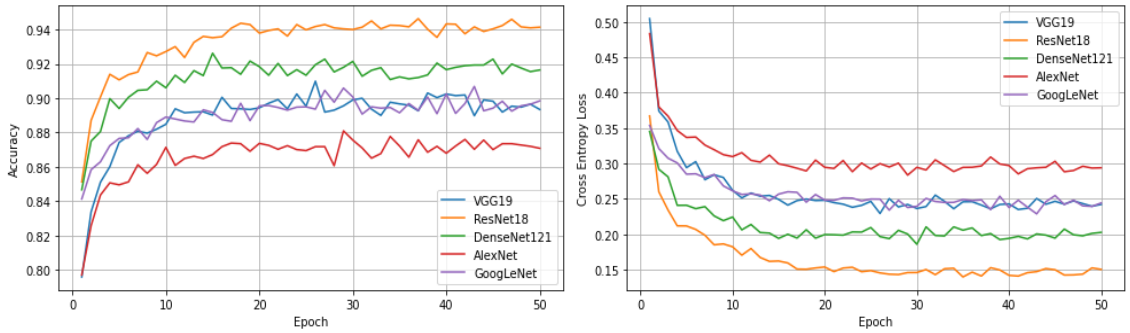

Figura 3. Acurácia e *loss* ao longo das épocas de cada um dos modelos do agrupamento. Os dados são do conjunto de treinamento.

O modelo professor usa a combinação das saídas dos modelos treinados para fazer as predições. As saídas desses modelos foram combinadas calculando a média das *logits* da última camada linear. Em seguida, as predições foram computadas usando o valor médio das *logits*. Ao fazer isso o agrupamento de redes obteve uma acurácia de 94.71%. Também foi experimentado fazer a predição por votação, utilizando a predição de cada modelo separadamente e escolhendo a classe com mais votos, contudo essa abordagem obteve uma acurácia de 94.55%. A Tabela 2 apresenta a acurácia de cada um dos modelos no conjunto de teste. As demais métricas do modelo professor (agrupamento de redes) são apresentadas na Tabela 3.

Tabela 2. Resumo da acurácia de cada um dos modelos do agrupamento no conjunto de teste.

Modelo	Sucesso na classificação (%)
VGG19	91.35
AlexNet	92.47
GoogLeNet	92.63
DenseNet121	93.11
ResNet18	94.39
Agrupamento de Redes (predição por votação)	94.55
Agrupamento de Redes (predição por média)	94.71

Tabela 3. Resumo da precisão, cobertura e F1-score para cada classe do modelo professor.

	Precisão	Cobertura	F1-score	Número de amostras
Normal	0.976	0.880	0.926	234
Pneumonia	0.932	0.987	0.959	390

Em seguida, foi feito o treinamento do modelo estudante. Inicialmente o modelo estudante foi treinado de maneira convencional. Este treinamento foi feito para que se comparasse o desempenho do modelo estudante treinado convencionalmente com desempenho do modelo treinado com a técnica de destilação do conhecimento obscuro. Após 25 épocas de treinamento, o modelo atingiu uma acurácia de 86.22% no conjunto de teste. As demais métricas são apresentadas na Tabela 4.

Tabela 4. Resumo da precisão, cobertura e F1-score para cada classe do modelo estudante treinado convencionalmente.

	Precisão	Cobertura	F1-score	Número de amostras
Normal	0.957	0.662	0.783	234
Pneumonia	0.829	0.982	0.899	390

Como discutido anteriormente, na seção 2.2, o uso da técnica de destilação de conhecimento obscuro introduz três novos hiper parâmetros: T, α e β. Nos experimentos desse trabalho utilizamos valores de T entre 2 e 5, sendo que os melhores resultados foram obtidos com T=2. Já para α e β, definiu-se $\beta = 1 - \alpha$ e foram experimentados valores de α=0.9, α=0.95 e α=0.99. Os melhores resultados foram obtidos com α=0.9. Ao treinar o modelo estudante com a técnica de destilação do conhecimento obscuro (T=2, α=0.9 e β=0.1), atingiu-se uma acurácia de 89.90% no conjunto de teste. As demais métricas são apresentadas na Tabela 5.

Tabela 5. Resumo da precisão, cobertura e F1-score para cada classe do modelo estudante treinado com a técnica de destilação do conhecimento obscuro.

	Precisão	Cobertura	F1-score	Número de amostras
Normal	0.948	0.774	0.852	234
Pneumonia	0.878	0.974	0.923	390

4.2. Comparação dos resultados do modelo aprendiz com os resultados do modelo professor e análise da eficácia da técnica de destilação do conhecimento

Baseado nesses resultados, percebe-se que é possível melhorar o desempenho de uma rede neural com a utilização da técnica de destilação do conhecimento obscuro. O modelo estudante treinado convencionalmente teve a acurácia 3.68% abaixo do modelo treinado com a destilação do conhecimento obscuro, constatando-se que é possível transferir conhecimento de uma rede complexa para um modelo mais simples.

Em bases de dados desbalanceadas (como a usada nesse trabalho) a rede tende a fazer mais predições da classe dominante (a classe com mais amostras de treinamento). Analisando as tabelas 3, 4 e 5, podemos perceber que o desbalanceamento da base de dados pode ter impactado no aprendizado do modelo. Ao compararmos as tabelas 4 e 5, podemos perceber que com o uso da destilação do conhecimento obscuro o modelo fez menos predições da classe 1 (pneumonia, que é a classe dominante), indicando que uso da técnica reduziu os efeitos do desbalanceamento da base de dados.

Além disso, é possível tornar o treinamento mais rápido, uma vez que com o uso da técnica de destilação do conhecimento obscuro foi possível atingir um desempenho melhor com um número menor de épocas, isso pode ser melhor visualizado na Figura 4, que apresenta os gráficos de acurácia e *loss* ao longo do treinamento. Também é importante destacar a possibilidade de embarcar o modelo estudante, visto que sua arquitetura mais simples viabiliza a utilização do mesmo em dispositivos com processamento limitado, sendo essa uma vantagem em se utilizar o modelo estudante ao invés de modelos mais complexos.

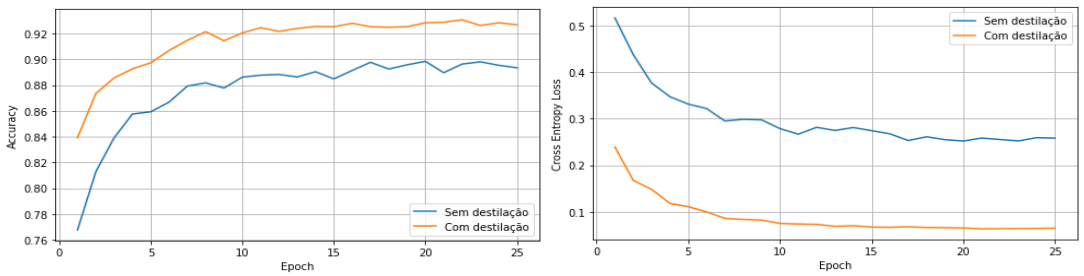

Figura 4. Acurácia e *loss* ao longo das épocas com e sem destilação do conhecimento. Os dados são do conjunto de treinamento.

4.3. Comparação dos resultados obtidos com outros trabalhos da literatura

Outros trabalhos da literatura já utilizaram modelos previamente treinados com a base de dados ImageNet [Russakovsky 2015] na tarefa de detecção de pneumonia a partir de imagens de raio X do tórax. Alguns utilizando a mesma base de dados (e as mesmas partições de treinamento e teste) que as usadas neste trabalho: Kermany et al. (2018) conseguiu 92.8% de acurácia aproveitando os pesos de uma rede Inception v3 [Szegedy et al. 2016], Chouhan et al. (2020) fez experimentos com cinco diferentes modelos previamente treinados: AlexNet [Krizhevsky, Sutskever e Hinton 2012], DenseNet [Huang et al. 2017], ResNet18 [He et al. 2016], Inception V3 [Szegedy et al. 2016] e GoogleNet [Szegedy et al. 2015], além de combinar as saídas dos cinco modelos anteriores (agrupamento de redes neurais), onde obteve o melhor desempenho (96.39% de acurácia). Em ambos os trabalhos citados, os resultados foram superiores aos obtidos nesse artigo, ao custo de maior esforço computacional se comparado ao modelo estudante. A Tabela 6 apresenta a comparação dos resultados com os trabalhos de Kermany et al. (2018) e de Chouhan et al. (2020).

Tabela 6. Quadro comparativo geral.

Modelo	Sucesso na classificação (%)
Modelo Estudante (treinamento convencional)	**86.22**
Modelo estudante (destilação do conhecimento obscuro)	**89.90**
Densenet121 [Chouhan et al. 2020]	92.62
Inception V3 [Kermany et al 2018]	92.80
AlexNet [Chouhan et al. 2020]	92.86
GoogLeNet[Chouhan et al. 2020]	93.12
Resnet18 [Chouhan et al. 2020]	94.23
Modelo Professor (agrupamento de redes)	**94.71**
Agrupamento de redes [Chouhan et al. 2020]	96.39

Uma boa maneira de mensurar a complexidade de um modelo é analisando a quantidade de parâmetros do mesmo, a Tabela 7 apresenta um comparativo da quantidade de parâmetros de cada um dos modelos. A quantidade de parâmetros do agrupamento de redes é no mínimo o somatório dos parâmetros dos modelos que o compõem. Destaca-se a quantidade de parâmetros consideravelmente menor do modelo estudante em relação aos outros modelos apresentados.

Tabela 7. Comparação da quantidade de parâmetros de cada modelo.

Modelo	Quantidade de parâmetros
Modelo estudante	**30778**
GoogLeNet	5601954
DenseNet121	6955906
ResNet18	11177538
Inception V3	24348900
AlexNet	57012034
VGG19	139589442
Modelo Professor (Estimativa)	244.685.774

5. Considerações Finais

Nesse artigo, é proposta uma abordagem baseada em redes neurais convolucionais para detecção de pneumonia com base em imagens de raio X do tórax. A abordagem focou no uso da técnica de destilação do conhecimento obscuro, com o objetivo de aproximar o desempenho de um modelo simples ao de um modelo complexo. Os resultados obtidos pelos modelos propostos indicaram a eficácia da técnica, uma vez que o uso da mesma melhorou o desempenho do modelo. Desse modo, a destilação do conhecimento obscuro se mostrou útil para a aplicação escolhida nos experimentos. Os resultados também demonstram que o modelo proposto no artigo pode ser uma alternativa em situações com limitações de hardware, uma vez que possui um desempenho razoável e um custo computacional menor que outros modelos da literatura.

Trabalhos futuros podem tentar melhorar o desempenho do modelo professor, por exemplo com o uso de um agrupamento de redes neurais feito por Chouhan et al. (2020), de modo que mais conhecimento seja transferido ao modelo estudante. A arquitetura do modelo estudante também pode ser aprimorada, isso pode ser feito com a adição de atalhos entre as camadas convolucionais, como as existentes na ResNet [He et al. 2016]. Além disso, a base de dados pode ser trabalhada sem a junção das classes de pneumonia viral e bacteriana, visto que estas apresentam significativas diferenças clínicas que impactam no diagnóstico diferencial. Por fim, podem ser experimentados mais valores para T, α e β.

Referências

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. In Advances in neural information processing systems (pp. 2654-2662).

Chouhan, V. et al. (2020). A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images. Applied Sciences, 10(2), 559.

Franquet, T. (2001). Imaging of pneumonia: trends and algorithms. European Respiratory Journal, 18(1), 196-208.

Gong, W., Chen, H., Zhang, Z., Zhang, M., Wang, R., Guan, C., & Wang, Q. (2019). A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion. Sensors, 19(7), 1693.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Dong, J. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122-1131.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Längkvist, M., Karlsson, L., and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42:11– 24.

Rajpurkar, P. et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

Russakovsky, O. et al. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

Szegedy, C. et al. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Tang, Z., Wang, D., & Zhang, Z. (2016, March). Recurrent neural network training with dark knowledge transfer. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5900-5904). IEEE.

World Health Organization. (2001). Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children (No. WHO/V&B/01.35). World Health Organization.