Extensão da base de dados RYDLS-20 para pesquisas em identificação de COVID-19 em imagens de raio-X

Luiz Fellipe Machi Pereira (PIBIC/CNPq/FA/Uem), Yandre Maldonado e Gomes da Costa (Orientador), Lucas de Oliveira Teixeira (Coorientador), e-mails: ra103491@uem.br, yandre@din.uem.br, loteixeira2@din.uem.br.
Universidade Estadual de Maringá / Centro de Tecnologia, PR.

Ciência da Computação/ Metodologia e Técnicas da Computação

Palavras-chave: base de dados de raio-x, extensão de base de dados, identificação de COVID-19.

Resumo:

Este trabalho teve como objetivo a extensão de uma base de dados com imagens de raio-x de tórax, a fim de que se possa utilizar esses dados em experimentos para a identificação de COVID-19. Neste trabalho incrementamos a base de dados RYDLS-20, disponibilizada gratuitamente com o objetivo de apoiar o desenvolvimento de tarefas de identificação de doenças em radiografias torácicas. São apresentados experimentos básicos realizados na versão original e na versão estendida, explorando algumas de suas possíveis tarefas de identificação, como a identificação de COVID-19. Duas abordagens foram testadas, ambas utilizando redes neurais para extração das características, bem como o processo de identificação. Os resultados alcançados mostraram que o processo de extensão e balanceamento da nova base proporcionam um aumento do desempenho em alguns cenários. As melhores pontuações para a métrica F1-Score foram 0,95 no conjunto de validação e 0,94 no conjunto de teste, para a classe de COVID-19 na identificação dessa doença, utilizando uma rede neural com base no modelo ResNet-50.

Introdução

A pandemia provocada pelo avanço global do novo coronavírus tem causado sérios impactos econômicos e sanitários no mundo todo. A doença provocada pelo vírus SARS-CoV-2, oficialmente denominada como COVID-19, tem se alastrado de forma preocupante e, muito por isso, vem chamando a atenção de toda a comunidade científica, de diferentes áreas do conhecimento, a fim de mover esforços para combater o problema com diferentes perspectivas.

O desenvolvimento de sistemas de classificação para o reconhecimento de padrões em diferentes problemas, depende de forma geral da disponibilidade de bases de dados bastante robustas com amostras dos padrões que se pretende classificar. Em linhas gerais, quanto maior a base em termos de quantidade de amostras, maior tende a ser a taxa de acerto que se consegue alcançar como resultado final do sistema criado. Em um

trabalho recente, os autores organizaram um conjunto de imagens de raio-X que inclui amostras de pulmões afetados por diferentes tipos de pneumonia, incluindo COVID-19. O resultado foi uma base de dados disponível publicamente para a comunidade de pesquisa, chamada RYDLS-20 (Pereira, 2020). Diante disso, o objetivo deste trabalho é incrementar e expandir a base RYDLS-20 com mais imagens e mais tipos de pneumonia causados por diferentes patógenos.

Materiais e métodos

A base de dados: a base de dados RYDSL-20 original possuía inicialmente 1144 imagens de radiografias torácicas, advindas de diferentes fontes e com diferentes tamanhos, sendo que 1000 dessas imagens eram referentes a radiografias de pacientes que não apresentavam nenhuma patologia, aqui chamadas de radiografias normais, e apenas 90 imagens de pacientes com COVID-19. A versão expandida criada conta com 58998 imagens de raio-x de tórax, sendo que 20644 são referentes a radiografias normais, 15567 apresentam alguma opacidade, em 11895 não se pode identificar o estado, 5788 são de pacientes com pneumonia, 899 de pacientes com tuberculose e 4205 imagens de pacientes com COVID-19. Para garantir que não houvesse imagens repetidas, criou-se um *hash* de imagens para cada classe, uma vez que só poderia existir uma imagem para cada entrada, desta forma foi possível identificar imagens duplicadas e removê-las da base de dados.
Classificadores: o cenário escolhido para a avaliação da base foi a classificação de COVID-19 por meio de imagens de radiografia de tórax. Para tanto, foram criadas quatro classes a partir dos rótulos das imagens base, a saber: “COVID-19”, referente aos pacientes que têm ou tiveram a doença; "Normal", a classe de pacientes com exame com resultado normal; "Pneumonia", referente aos pacientes com pneumonia causada por outros patógenos; e "Opacidade", a classe de pacientes com algum tipo de opacidade no pulmão, o que indicaria uma possível lesão no tecido pulmonar. Os modelos escolhidos para realizar a classificação foram uma rede neural com base no modelo ResNet-50, pois existem bons resultados apresentados na literatura (Nayak,2020; Narin, 2020), e uma rede neural com base no modelo Xception, por já ter sido usado em diferentes experimentos com a COVID-19 (Makris,2020).
Em ambas as redes foram adicionadas camadas após o modelo base. Uma representação geral das redes pode ser vista na Figura 1. As redes neurais foram treinadas utilizando 4205 imagens de cada classe, totalizando 16820 imagens, por 80 épocas e tamanho de lote igual a 25.

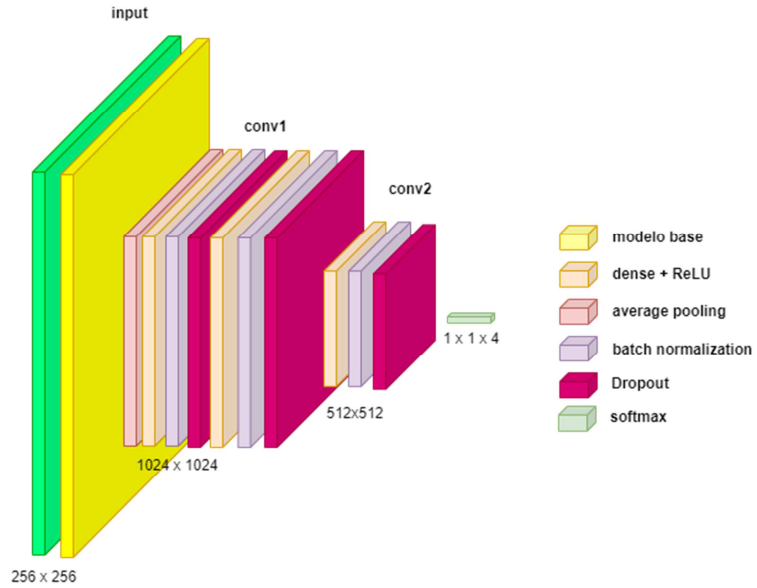

Figura 1 – Esquema geral das redes neurais criadas. Fonte: o autor.

Resultados e Discussão

Para analisar o desempenho dos resultados experimentais, foi escolhida a medida F1-Score. Os resultados para ambas as métricas e para ambos os conjuntos alvos (teste e validação) foram sumarizados na Tabela 1.
A classe que obteve melhores resultados para a métrica F1-Score, em ambos os conjuntos, foi a classe de pneumonia, chegando a 0,98 no conjunto de teste usando o modelo com ResNet-50, seguido pela classe de paciente com COVID-19, chegando a 0,95 no conjunto de teste com o mesmo modelo. É possível notar que, em geral, a rede neural com base no modelo ResNet-50 obteve melhores resultados para todas as classes para a métrica F1-Score.

Tabela 1 – F1-Score para os modelos testados nos conjuntos de validação e teste.

	Validação		**Teste**	
Classe	**ResNet-50**	**Xception**	**ResNet-50**	**Xception**
Normal	0,90	0,85	0,90	0,86
Opacidade	0,90	0,85	0,90	0,87
Pneumonia	0,97	0,96	0,98	0,96
COVID-19	0,94	0,85	0,95	0,87

Conclusões

Neste trabalho, apresentamos a extensão do banco de dados RYDLS-20, um banco de dados disponibilizado gratuitamente para a comunidade de pesquisa, que visa apoiar o desenvolvimento de tarefas de identificação de COVID-19 em imagens de raio-X de tórax. O banco de dados original é composto por 1144 imagens, sendo 90 de pacientes com COVID-19, 54 de pacientes com alguma opacidade no pulmão e 1000 de pacientes com nenhum dos problemas anteriores. Na nova versão da base foram incluídas 57854 imagens, totalizando 4205 imagens referentes a classe COVID-19, 20644 referentes a classe normal e 15567 referente a classe opacidade.
Também foram apresentados experimentos com a identificação de COVID-19 e seus resultados. Nos experimentos realizados, as características foram extraídas e classificadas por meio de modelos de rede neural, com a ResNet-50 e com Xception, sendo que, o modelo ResNet-50 obteve melhores resultados para todas as classes para a métrica F1-Score.
No futuro, pretendemos expandir a base com amostras de tomografias computadorizada e também a inclusão de radiografias de outras doenças. Pretendemos também realizar novos experimentos, desta vez utilizando classificação hierárquica a fim de comparar os resultados e aumentar as taxas de classificação.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro na concessão da bolsa, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e aos orientadores pelo suporte durante o desenvolvimento deste trabalho.

Referências

Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla Jr, C. N., & Costa, Y. M. (2020). **COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios**. *Computer Methods and Programs in Biomedicine*, *194*, 105532.
Nayak, S. R., Nayak, D. R., Sinha, U., Arora, V., & Pachori, R. B. (2021). **Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study**. B*iomedical Signal Processing and Control*, 64, 102365.
Narin, A., Kaya, C., & Pamuk, Z. (2021). **Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks**. *Pattern Analysis and Applications*, 1-14.
Ko, H., Chung, H., Kang, W. S., Kim, K. W., Shin, Y., Kang, S. J., ... & Lee, J. (2020). **COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation**. *Journal of medical Internet research*, *22*(6), e19569.