A Preliminary Exploration into the Performance of Severity Encoding Strategies for Deep Learning-Based Severity Stratification of COVID-19 Patients using Chest X-Rays on A Clinical Site Cohort

Alexander MacLean — Vision and Image Processing Lab, University of Waterloo
Jackson Zheng — Vision and Image Processing Lab, University of Waterloo
Tia Tuinstra — Vision and Image Processing Lab, University of Waterloo
Beiyi Shen — Stony Brook School of Medicine, Department of Radiology
Almas Abbasi — Stony Brook School of Medicine, Department of Radiology
Mahsa Hoshmand Kochi — Stony Brook School of Medicine, Department of Radiology
Tim Q. Duong — Montefiore Medical Center and Albert Einstein College of Medicine
Alexander Wong — Vision and Image Processing Lab, University of Waterloo
Email: {alex.maclean, j75zheng, ttuinstra, a28wong}@uwaterloo.ca, tim.duong@einsteinmed.org

Abstract

A critical step in the clinical workflow for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) patients is lung disease severity assessment, providing valuable information to aid in effective patient care and management as well as treatment planning. Given the difficulty of performing such assessments by healthcare workers and the necessity of expert radiologists who are already burdened by the significant load caused by the pandemic, one promising direction is the use of computer-aided decision support systems powered by deep learning. An important design consideration in the building of deep neural networks for SARS-CoV-2 disease severity assessment is in the way severity scores are encoded, as it can have a big influence on both the training and inference aspects of the neural network. In this study, we explore the performance impact of different severity encoding strategies for deep learning-based severity stratification of COVID-19 patients using chest x-rays (CXRs) on a clinical site cohort collected from the Stony Brook University Hospital. More specifically, we study the impact of different quantized severity encoding schemes, different granularity in the severity encoding, as well as compare quantized encoding vs. continuous encoding vs. hybrid centroid weighted encoding.

1 Introduction

During the on-going Coronavirus Disease 2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), significant focus has gone into exploring the efficacy of different technology advances for aiding clinicians in combating the disease. One particular area of interest in this technological exploration involves exploring the efficacy of machine learning techniques such as deep learning to aid clinicians in the clinical decision support process. This is of particular importance given the significant burden on clinicians, radiologists, and healthcare workers due to the surge in patient intake, and the potential for machine learning to reduce that burden.

Significant progress and advancements have been made in the area of deep learning-driven COVID-19 detection using different imaging modalities such as chest x-rays (CXR) [1], computed tomography (CT) scans [2], and point-of-care ultrasound (POCUS) [3]. In addition to COVID-19 detection, it is also important for clinicians to be able to identify the severity of infections in SAR-CoV-2 positive patients to direct proper treatment and allocation of resources. One severity assessment strategy identified in literature involves radiologists studying CXR images and assigning severity scores based on: a) the density of opacities seen in the lungs (opacity extent), and b) the geographic spread of such opacities across the lungs (geographic extent) [4, 5]. In this system, geographic extent and opacity extent are each scored on a scale of 0-4 for each individual lung, and the scores for each metric are summed across both lungs in an image, resulting in one score on a scale of 0-8 for geographic extent and one for opacity extent.

One challenge with severity assessment, even with the help of the aforementioned severity assessment strategies, is that it is very difficulty for healthcare workers to interpret CXR images given the subtleties of the condition at different severity levels. As such, currently such assessments often necessitate expert radiologists, who are already burdened by the significant load caused by the pandemic. Therefore, computer-aided severity assessment powered by deep learning can have significant benefits to aid clinicians and is highly desired, with promising results shown in a number of recent studies [6, 7].

An important design consideration in the building of deep neural networks for SARS-CoV-2 disease severity assessment lies in the manner in which severity scores are encoded. The encoding scheme used can have a significant impact on both the training behaviour of the neural network, as well as during inference time when the neural network is making a prediction of lung disease severity. In this study, we investigate the impact of different severity encoding strategies on the severity stratification performance of COVID-19 patients using chest x-rays (CXRs) on a clinical site cohort collected from the Stony Brook University Hospital.

2 Methodology

For this study, a clinical site cohort collected from the Stony Brook University Hospital was used. Ethics clearance was received from the Stony Brook University Institution Review Board Office of Research Compliance research ethics board. The cohort consists of CXR images from 2372 COVID-positive patient cases, split into training, validation, and testing sets of 1518, 380, and 475 images, respectively. Radiological scoring was performed by two board-certified chest radiologists with 20+ years of experience (A.A. and M.H.) and a 2nd-year radiology resident (B.S.). Each image was scored for both geographic extent and opacity extent as previously defined on a scale of 0.0 to 8.0, and the average score from the two radiologists for each metric is used in this study. Figure 1 shows 4 example CXR images from the cohort, with patient cases exhibiting different levels of geographic extent and opacity extent.

Exploration of the cohort, as shown in Figure 2, reveals that the distributions vary significantly across the different severity scores in question. The geographic extent scores are heavily left-skewed, with the upper end of the 0-8 scale having a significantly higher density of patient cases, while the opacity extent scores are normally distributed around the central value of 4 with very few patient cases being found at either extreme of the scoring scale. This difference in distribution suggests that being able to predict geographic extent and opacity separately is important for clinical use cases. Additionally, it should be noted that the number of scores falling on whole numbers (e.g., 1.0, 2.0, etc.) is generally higher than scores falling in between that occur when averaging two radiologist scores (e.g. 0.5, 1.5, 2.5, etc.) but the trends are still very clearly left-skewed and normal respectively.

To conduct this study, we leveraged the COVID-Net CXR-2 deep convolutional neural network architecture [8] as the backbone architecture for the various severity assessment deep neural networks that were constructed for severity prediction based on the different severity encoding strategies tested. COVID-Net CXR-2 was leveraged, as it was demonstrated empirically to provide strong discriminative latent representations of SARS-CoV-2 infection characteristics [8], and thus acts as a good foundation for exploring the efficacy of different encoding strategies.

The following experiments were conducted to allow us to better understand the different aspects of severity encoding. First, we investigate the influence of one-hot quantized severity encoding and integer quantized severity encoding on categorical severity prediction performance. In this experiment, we leveraged a three-level quantization encoding scheme, where severity scores from 0-8 were quantized into three levels: 0-3, 3-6, and 6-8.

Second, we investigate the impact of quantization level granularity on categorical severity prediction performance when leverag-

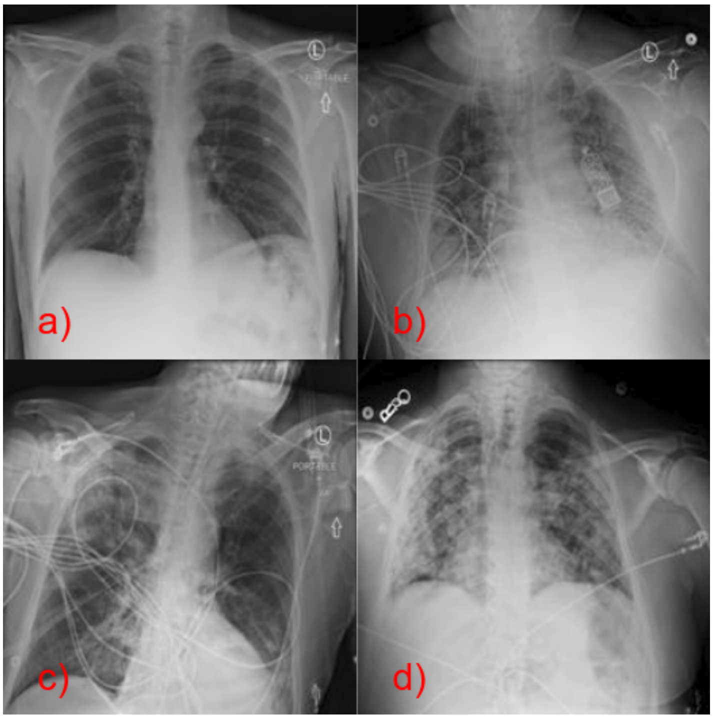

Fig. 1: Example CXR images from the Stony Brook cohort; a) low opacity density and geographic extent, b) higher geographic extent than opacity density, c) higher opacity density than geographic extent, and d) high geographic extent and high opacity density.

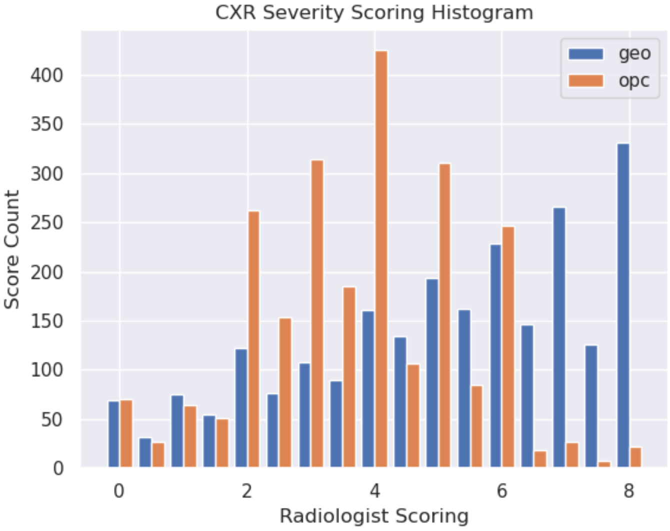

Fig. 2: Distribution of geographic extent and opacity extent scores averaged across the radiologists.

ing quantized severity encoding. This was accomplished by varying the quantization levels used in the quantized encoding ranging from three-level encoding to five-level encoding.

Third, we investigate the impact of distribution-driven loss weighting during training on categorical severity prediction performance using quantized severity encoding. The weights were determined based on the data distribution for each quantized severity level using Eq. 1:

$$weight = \frac{n}{n_c * n_j} \tag{1}$$

Where n is the total amount of patient cases in the training dataset, n_c is the number of severity levels, and n_j is the number of patient cases for the j^{th} level. Using the three-level quantized encoding as an example, levels 0, 1, and 2 represent low, medium, and high severity respectively. Sensitivity and accuracy were leveraged for quantitative evaluation for the first three experiments.

Fourth and finally, we compare the performance of continuous severity encoding and quantized severity encoding schemes for continuous severity prediction. More specifically, for quantized severity encoding, we evaluated using different severity level granularities ranging from three level encoding to five level encoding, with the weighted average of the quantized severity centroids for each quantized level, weighted by the softmax output for the corresponding quantized level. This is compared to continuous severity encoding, where the severity score is used directly as a continous value and the model trained using a mean-squared error (MSE) loss.

3 Results and Discussion

3.1 Quantized Severity Encoding: One-Hot Encoding vs. Integer Encoding

Table 1 compares the categorical severity prediction results for deep convolutional neural networks trained using three-level, equal-weighted one-hot quantized severity encoding and integer quantized severity encoding for the geographic extent scores. It can be observed that the accuracy of the network trained using one-hot encoding was noticeably higher than that trained using integer encoding. Therefore, for the rest of the experiments we focus on one-hot encoding for quantized severity encoding comparisons.

Encoding	Sensitivity (level 0)	Sensitivity (level 1)	Sensitivity (level 2)	Accuracy
One-Hot	0.686	0.603	0.827	0.740
Integer	0.629	0.618	0.827	0.705

Table 1: Categorical geographic extent prediction performance for one-hot and integer quantized severity encoding with quantized level mappings of (0,1,2)->[0-3, 3-6, 6-8].

3.2 Quantization Level Granularity

Table 2, Table 3, and Table 4 depict the categorical prediction performance for geographic extent and opacity extent using three-level, four-level, and five-level quantized severity encoding, respectively.

Metric	Sensitivity (levels 0 - 2)			Accuracy
Geo	0.686	0.603	0.827	0.740
Opc	0.598	0.745	0.644	0.631

Table 2: Categorical prediction performance for geographic and opacity extent using three-level quantized severity encoding

Metric	Sensitivity (levels 0 - 3)				Accuracy
Geo	0.727	0.420	0.535	0.760	0.637
Opc	0.600	0.804	0.459	0.422	0.496

Table 3: Categorical prediction performance for geographic and opacity extent using four-level quantized severity encoding

Metric	Sensitivity (levels 0 - 4)					Accuracy
Geo	0.667	0.564	0.278	0.612	0.600	0.557
Opc	0.667	0.556	0.355	0.500	0.375	0.482

Table 4: Categorical prediction performance for geographic and opacity extent using five-level quantized severity encoding

It can be observed that as the granularity of the quantized severity encoding increased, the sensitivity and accuracy decreased. More specifically, it was observed that there was a progressive increase in false negatives being attributed to neighbouring severity levels as granularity increased. That said, while coarser granularity lead to higher accuracy and sensitivity, it may potentially be less informative from a clinical interpretation perspective when using this information to inform patient care and treatment planning. Therefore, this trade-off between accuracy and actionability needs to be discussed further with clinicians in order to understand which factors are most important to aid them in their workflows to guide future work.

3.3 Distribution-driven Loss Weighting

Table 5 and Table 6 shows the categorical severity prediction performance results of leveraging distribution-driven loss weighting during training when using the three-level and four-level quantized severity encoding, respectively.

Metric	Sensitivity (levels 0 - 2)			Accuracy
Geo	0.700	0.695	0.721	0.708
Opc	0.815	0.457	0.533	0.555

Table 5: Categorical predictive performance for geographic and opacity extent using three-level quantized severity encoding using distribution-driven loss weighting of [1.809, 0.966, 0.707] and [1.3764, 0.541, 2.345], respectively

Metric	Sensitivity (levels 0 - 3)				Accuracy
Geo	0.788	0.362	0.616	0.648	0.600
Opc	0.833	0.642	0.293	0.578	0.494

Table 6: Categorical predictive performance for geographic and opacity extent using four-level quantized severity encoding using distribution-driven loss weighting of [2.878, 1.376, 0.959, 0.530] and [3.16, 0.641, 0.605, 2.111] respectively

The addition of distribution-driven loss weighting did not lead to significant improvement in prediction performance. Underrepresented severity levels, which differed between metrics as shown in Figure 2, showed better sensitivity, while more data-rich severity levels suffered a decrease in performance, which represents a trade-off that needs to be discussed.

3.4 Continuous Severity Prediction Performance

Table 7 show the performance of continuous severity encoding and quantized severity encoding schemes for continuous severity prediction of geographic extent. Figure 3 displays the corresponding results in scatter-plot form.

Model	R^2
Continuous encoding	0.676
Three-level quantized encoding	0.654
Four-level quantized encoding	0.670
Five-level quantized encoding	0.648

Table 7: Continuous severity prediction performance for geographic extent for continuous encoding and quantized encoding schemes

Table 8 show the performance of continuous severity encoding and quantized severity encoding schemes for continuous severity prediction of opacity extent. Figure 4 displays the corresponding results in scatter-plot form.

Model	R^2
Continuous encoding	0.508
Three-level quantized encoding	0.395
Four-level quantized encoding	0.477
Five-level quantized encoding	0.502

Table 8: Continuous severity prediction performance for opacity extent for continuous encoding and quantized encoding schemes

The results seen when comparing continuous severity prediction performance for continuous encoding and quantized encoding schemes are very interesting. Despite the quantized severity encoding schemes not explicitly using the continous severity score directly via regression training, some of the quantized encoding schemes led to similar performance as when continuous encoding was leveraged directly.

Additionally, even though increasing quantization granularity in the quantized severity encoding led to a decrease in accuracy for categorical severity prediction, this trend was not observed in the case of continuous severity prediction. This suggests there is merit in examining if the increased encoding granularity could provide better representation that could improve outcomes in the continuous severity prediction scenario.

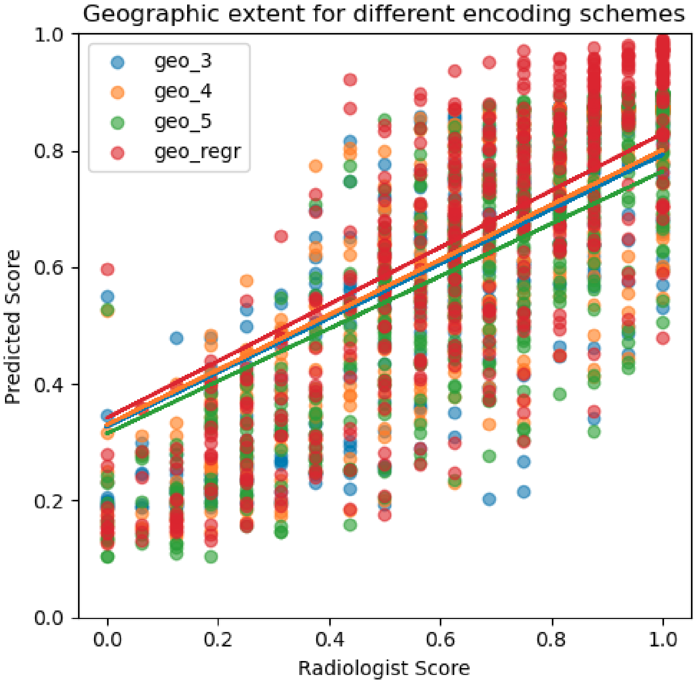

Fig. 3: Geographic extent for different encoding schemes.

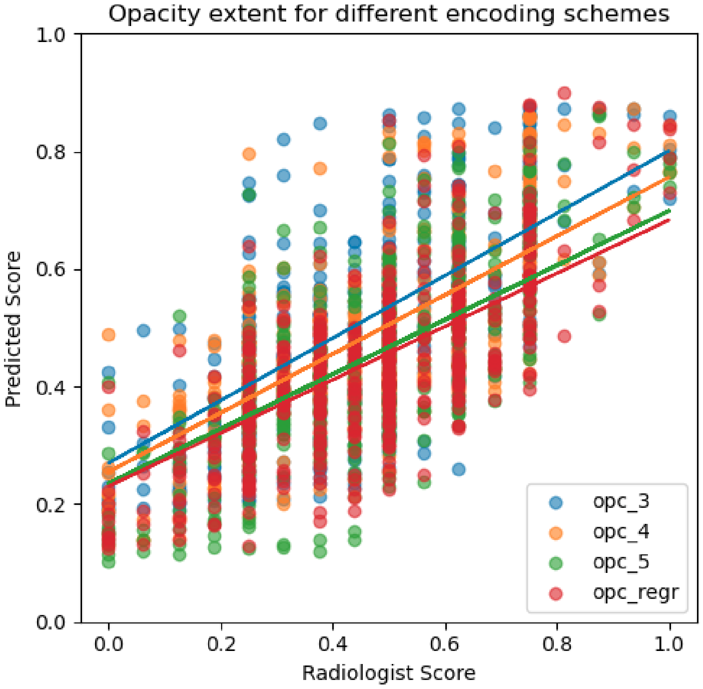

Fig. 4: Opacity extent for different encoding schemes.

Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs program. The study has received ethics clearance from the University of Waterloo (42235). All experimental protocols were approved by University of Waterloo. All methods were carried out in accordance with University of Waterloo ethics guidelines and regulations. This was a retrospective study approved by the Stony Brook University Institution Review Board Office of Research Compliance, approval number IRB2020-00207, and was exempt from informed consent and Health Insurance Portability & Accountability Act (HIPAA) waiver.

References

[1] L. Wang, Z. Q. Lin, and A. Wong, “COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images,” *Scientific Reports*, vol. 10, no. 1, Nov. 2020.

[2] H. Gunraj, L. Wang, and A. Wong, “COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images,” *Frontiers in Medicine*, vol. 7, Dec. 2020.

[3] A. MacLean, S. Abbasi, A. Ebadi, A. Zhao, M. Pavlova, H. Gunraj, P. Xi, S. Kohli, and A. Wong, “Covid-net us: A tailored, highly efficient, self-attention deep convolutional neural network design for detection of covid-19 patient cases from point-of-care ultrasound imaging,” 2021.

[4] H. Y. F. Wong, H. Y. S. Lam, A. H.-T. Fong, S. T. Leung, T. W.-Y. Chin, C. S. Y. Lo, M. M.-S. Lui, J. C. Y. Lee, K. W.-H. Chiu, T. W.-H. Chung, E. Y. P. Lee, E. Y. F. Wan, I. F. N. Hung, T. P. W. Lam, M. D. Kuo, and M.-Y. Ng, “Frequency and distribution of chest radiographic findings in patients positive for COVID-19,” *Radiology*, vol. 296, no. 2, pp. E72–E78, Aug. 2020. [Online]. Available: https://doi.org/10.1148/radiol.2020201160

[5] M. A. Warren, Z. Zhao, T. Koyama, J. A. Bastarache, C. M. Shaver, M. W. Semler, T. W. Rice, M. A. Matthay, C. S. Calfee, and L. B. Ware, “Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS,” *Thorax*, vol. 73, no. 9, pp. 840–846, Jun. 2018. [Online]. Available: https://doi.org/10.1136/thoraxjnl-2017-211280

[6] A. Wong, Z. Q. Lin, L. Wang, A. G. Chung, B. Shen, A. Abbasi, M. Hoshmand-Kochi, and T. Q. Duong, “Covid-net s: Towards computer-aided severity assessment via training and validation of deep neural networks for geographic extent and opacity extent scoring of chest x-rays for sars-cov-2 lung disease severity,” 2021.

[7] H. Aboutalebi, M. Pavlova, M. J. Shafiee, A. Sabri, A. Alaref, and A. Wong, “Covid-net cxr-s: Deep convolutional neural network for severity assessment of covid-19 cases from chest x-ray images,” 2021.

[8] M. Pavlova, N. Terhljan, A. G. Chung, A. Zhao, S. Surana, H. Aboutalebi, H. Gunraj, A. Sabri, A. Alaref, and A. Wong, “Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images,” 2021.