World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:2, 2021

Bayesian Deep Learning Algorithms for Classifying COVID-19 Images

I. Oloyede

***Abstract*—**The study investigates the accuracy and loss of deep learning algorithms with the set of coronavirus (COVID-19) images dataset by comparing Bayesian convolutional neural network and traditional convolutional neural network in low dimensional dataset. 50 sets of X-ray images out of which 25 were COVID-19 and the remaining 20 were normal, twenty images were set as training while five were set as validation that were used to ascertained the accuracy of the model. The study found out that Bayesian convolution neural network outperformed conventional neural network at low dimensional dataset that could have exhibited under fitting. The study therefore recommended Bayesian Convolutional neural network (BCNN) for android apps in computer vision for image detection.

***Keywords*—**BCNN, CNN, Images, COVID-19, Deep Learning.

I. Introduction

DUE to recent development in computer vision algorithm and more powerful languages evolved such as Tensorflow developed by Google in [1] which makes both structured and unstructured analytics possible. Reference [2] developed active learning framework for high dimensional data which often posed difficulties in decades, thus demonstrated it on images classifications adopting BCNN.

Active learning, being a process of model learning that accommodates small amount of data which capture its uncertainty over the unseen data, was fully implemented and proved to be more efficient [2]. This study therefore adopts active learning since scanty images of COVID-19 patients are available, deep learning had been proved to outperformed traditional model based images classification, recognition, natural language processing and reinforcement/transfer learning which depend on complex assumption [3].

Deep learning required plenty data samples for high precision and accuracy [4], this shortcoming had been taken care of through Bayesian Deep Learning which adopts active learning with which small data are sufficient to perform accurately with high precision. Bayesian Deep Learning used deep neural network to ascertain uncertainty quantification of the output [5]. Reference [6] claimed that Bayesian Deep Learning is robust to overfitting and often influences proper decision making. They added that sampling on large scale images classification problems is feasible. They submitted that sampling posterior in multi-chain brought about accuracy of the posterior prediction and better uncertainty management.

Reference [7] opined that some images have consistent prediction while some have more than one classification results. Bayesian deep learning with variational inference captures out-of-distribution sample which conventional machine and deep learnings could not proffer solution to. It thus minimizes the uncertainty, with confidence score which derived from predictive variance of the model. This becomes easier with the development and deployment of Tensorflow probability into python programming computation. Reference [8] adopted Bayesian Deep Learning for data augmentation in images classification and claimed that their technique depicted better performance compared to traditional approach. Reference [9] opined that classification system of relic in choroidal optical coherence tomography angiography (OCTA) portend similar attributes for clinical interpretation. This is paramount in OCTA knowledge acquisition.

A classification system of these artifacts facilitates a systematic approach to image interpretation and serves as a bench mark for image grading in clinical trials that examines retinal images as an endpoint. Reference [11] examined Bayesian deep learning in a model-based interpretable approach nonlinear theory and its applications.

The sections of the study are arranged as follow: Section I contains introduction while Section II examined methodical design that thoughtfully synthesizes the statistical framework of the study. Data analysis and interpretation were set out in Section III while Section IV deals with conclusion. This study therefore compared BCNN and Convolutional neural network on the images of both normal and coronavirus (COVID-19) patients in a machine learning classification paradigm.

II. Methodical Design

A neural network in a probabilistic model paradigm with categorical dependent variable having bi-classes of 0 and 1, given dataset $D = \{x^{(i)}, y^{(i)}\}$, the likelihood function $p(D|w) = \prod_i p(y^{(i)}|x^{(i)}, w)$ has a function of parameters w. Maximizing the $p(D|w)$ gives the maximum likelihood estimate (MLE) of parameter $\boldsymbol{w}$. Multiplying the likelihood $p(D|w)$ with a prior distribution $p(\boldsymbol{w})$ will be proportional to the posterior density $p(\boldsymbol{w}|D) \propto p(D|\boldsymbol{w})p(\boldsymbol{w})$. Maximizing $p(D|\boldsymbol{w})p(\boldsymbol{w})$ gives the maximum a posteriori Probability (MAP) estimate of w. The estimation of MAP gives the mode of the posterior distribution with a regularizing effect that prevent overfitting. Thus optimization objectives of MLE with regularization term of log prior eliminate overfitting which ordinarily yields severe overfitting if MLE is adopted. This is because log prior that is added to MLE h regularization term eliminates overfitting [10].

Adopting posterior predictive distribution $p(y|x, D) =$

I. Oloyede is with the Department of Statistics, University of Ilorin, P.M.B 1515, Ilorin, Nigeria (phone: +2348053049890; e-mail: oloyede.i@unilorin.edu.ng)

$\int p(y|x,w)p(w|D)dw$ where nuisance parameters are hedged (marginalized) out thereby reduce the weight uncertainty that could have affected the accuracy and precision of the estimates.. This is like taken average of the predictions of weak and strong ensembles of neural networks divided by the posterior probabilities of the parameters w. Analytical solution for the posterior $p(w|D)$ in neural networks is intractable. Minimization of the Kullback-Leibler divergence between $q(w|\theta)$ and the true posterior $p(w|D)$ w.r.t. to θ is necessary to approximate the posterior density with a variational distribution $q(w|\theta)$ of the functional form of the model [10].

In line with [10] the Kullback-Leibler divergence between the variational distribution $q(\boldsymbol{w}|\boldsymbol{\theta})$ and posterior density $p(\boldsymbol{w}|D)$ is expressed as

$$KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) = \int q(\boldsymbol{w}|\boldsymbol{\theta}) \log \frac{q(\boldsymbol{w}|\boldsymbol{\theta})}{p(\boldsymbol{w}|D)} d\boldsymbol{w} \quad (1)$$

$$= Eq(\boldsymbol{w}|\boldsymbol{\theta}) \log \frac{q(\boldsymbol{w}|\boldsymbol{\theta})}{p(\boldsymbol{w}|D)} \quad (2)$$

Applying Bayes' rule to $p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})p(\boldsymbol{w})}{p(D)}$ we obtain

$$KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) = E_{q(\boldsymbol{w}|\boldsymbol{\theta})} log \frac{q(\boldsymbol{w}|\boldsymbol{\theta})}{p(D|\boldsymbol{w})p(\boldsymbol{w})} p(D) \quad (3)$$

Taking logarithm we have

$$= E_{q(\boldsymbol{w}|\boldsymbol{\theta})}[logq(\boldsymbol{w}|\boldsymbol{\theta}) - logp(D|\boldsymbol{w}) - logp(\boldsymbol{w}) + logp(D)] \quad (4)$$

$$= E_{q(\boldsymbol{w}|\boldsymbol{\theta})}[logq(\boldsymbol{w}|\boldsymbol{\theta}) - logp(D|\boldsymbol{w}) - logp(\boldsymbol{w})] + logp(D) \quad (5)$$

$$= KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, q(\boldsymbol{w})) - E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(D|\boldsymbol{w} + p(\boldsymbol{w})) + logp(D) \quad (6)$$

$E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(D|\boldsymbol{w} + p(\boldsymbol{w}))$, this is joint posterior density with regularization term(prior), its sum with variational term is the evidence lower bound that is expressed as: $E_{q(\boldsymbol{w}|\boldsymbol{\theta})}[logq(\boldsymbol{w}|\boldsymbol{\theta}) - logp(D|\boldsymbol{w}) - logp(\boldsymbol{w})]$, the normalizing constant $logp(D)$ does not depend on weight $\boldsymbol{w}$.

Let $y \in \{0,1\}$ be a binary random variable, the likelihood for a sequence of $\mathcal{D} = (x_1, \ldots, x_N)$ of COVID-19 image is

$$p(\mathcal{D}|w) = \prod_{n=1}^{N} w^{x_n}(1-w)^{1-x_n} \quad (7)$$

$$= w^{N_1}(1-w)^{N_0} \quad (8)$$

Let $\mathcal{D} = (x_1, \ldots, x_N)$ be set of data, the likelihood $l(w)$ can be expressed as (7) since the dependent variable is categorical, the Bernoulli distribution is appropriate, let $\in \{0,1\}$, given the $\mathcal{D} = (x_1, \ldots, x_N)$, the likelihood is:

$$p(\mathcal{D}|w) = \prod_{n=1}^{N} p(x_n|w) \quad (9)$$

$$= \prod_{n=1}^{N} w^{x_n}(1-w)^{1-x_n} \quad (10)$$

$$= w^{N_1}(1-w)^{N_0} \quad (11)$$

The prior is

$$p(\theta) = \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \quad (12)$$

The posterior density is

$$p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w) \quad (13)$$

$$\propto [w^{N_1}(1-w)^{N_0}][w^{\alpha_1 - 1}(1-w)^{\alpha_0 - 1}] \quad (14)$$

$$= w^{N_1 + \alpha_1 - 1}(1-w)^{N_0 + \alpha_0 - 1} \quad (15)$$

$$= KLq\big((\boldsymbol{w}|\boldsymbol{\theta})||q(\boldsymbol{w})\big) - E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(w^{N_1 + \alpha_1 - 1}(1-w)^{N_0 + \alpha_0 - 1}) + logp(D) \quad (16)$$

The normalizing constant $logp(D)$ is removed since it does not contribute to w, rearranging the model we have

$$KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) = E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(w^{N_1 + \alpha_1 - 1}(1-w)^{N_0 + \alpha_0 - 1}) \quad (17)$$

The RHS is referred to as *variational free energy* $F(D, \boldsymbol{\theta})$ and is also known as *evidence lower bound* $L(D, \boldsymbol{\theta})$ when it is non-negative and expressed as:

$$KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) = F(D, \boldsymbol{\theta}) \quad (18)$$

minimize $KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D))$ w.r.t. $\boldsymbol{\theta}$

$$KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) = -L(D, \boldsymbol{\theta}) \quad (19)$$

It is a lower bound on $logp(D)$ because the Kullback-Leibler (KL) divergence is always non-negative.

$$L(D, \boldsymbol{\theta}) L(D, \boldsymbol{\theta}) = logp(D) - KL(q(\boldsymbol{w}|\boldsymbol{\theta}) \,||\, p(\boldsymbol{w}|D)) \leq logp(D) \quad (20)$$

Thus the KL divergence between the variational distribution $q(\boldsymbol{w}|\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{w}|D)$ is also minimized by maximizing the evidence lower bound.

$$F(D, \boldsymbol{\theta}) = KL(q(\boldsymbol{w}|\boldsymbol{\theta})||p(\boldsymbol{w})) - E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(D|\boldsymbol{w}) \quad (21)$$

This refers to as variational free energy (VFE). The Kullback-Leibler Divergence (KLD) which optimizes variational distribution $\mathrm{q}(\mathbf{w}|\boldsymbol{\theta})$ and prior $\mathrm{p}(\mathbf{w})$ is known as Complexity Cost (CC) of the distribution. The expected value of log likelihood $logp(D|\boldsymbol{w})$ and variational distribution $\mathrm{q}(\mathbf{w}|\boldsymbol{\theta})$ is the Likelihood Cost (LC) [10]. The Cost Function can be expressed as

$$F(D, \boldsymbol{\theta}) = E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logq(\boldsymbol{w}|\boldsymbol{\theta}) - E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(\boldsymbol{w}) - E_{q(\boldsymbol{w}|\boldsymbol{\theta})} logp(D|\boldsymbol{w}) \quad (22)$$

The study observed that all three terms in (22) are expectations with respect to the variational distribution $q(\boldsymbol{w}|\boldsymbol{\theta})$. The cost function can therefore be approximated by drawing samples $\boldsymbol{w}^{(i)}$ from $q(\boldsymbol{w}|\boldsymbol{\theta})$

$$.F(D,\boldsymbol{\theta}) \approx 1N\sum i = \frac{1}{N}\sum_{i=1}^{N}[logq(\boldsymbol{w}^{(i)}|\boldsymbol{\theta}) - logp(\boldsymbol{w}^{(i)}) - logp(D|\boldsymbol{w}^{(i)})] \quad (23)$$

Data used for the study were obtained from [12].

III. Data Analysis and Interpretation

The number of kernels, dropping out rate, (Convolution2Dreparameterization, Flatten) and output units (DenseFlipout) are arbitrary selected, without any parameter tuning with 100 and 1000 batches, Bayesian approach implemented gradient which capture over/under-fitting. Python [12] and anaconda [13] were statistical tools used to analyze the data.

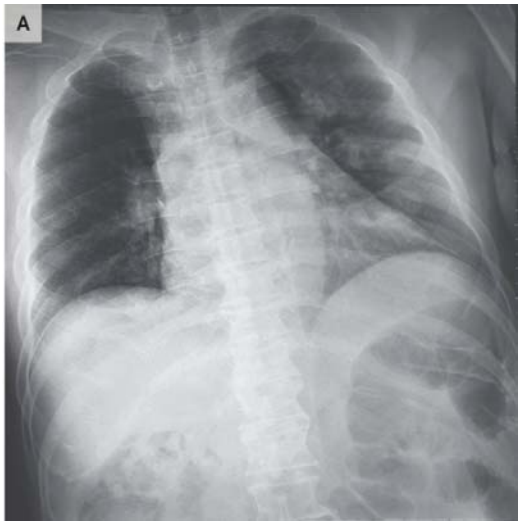

Fig. 1 Sample X-rays of COVID-19 patient

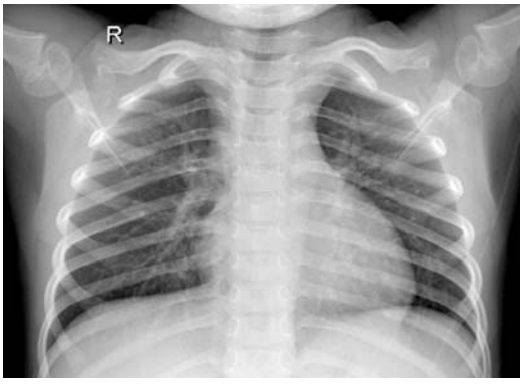

Fig. 2 Sample X-rays of a normal patient

The study compared CNN with Bayesian CNN on image classification of COVID-19 x-ray. This study implemented Flipout gradient estimator to minimize the negative evidence lower bound (ELBO) as the loss. It computes the integration when deriving the posterior distribution. Bayesian CNN accommodates uncertainty measure of the weights and predictions of which traditional CNN could not.

TABLE I
Performance Criteria of CNN and BCNN with Different Epoch

Iterations	1000		100	
	BNN	CNN	CNN	BNN
Validation Accuracy	0.9514	0.8758	0.8413	0.94
Training Loss	0.0048	0.0017	0.0353	2.0321
Validation Loss	0.1321	0.1591	0.4571	1.7408
Predicted Validation	6.294	6.6838	5.9685	59.899
Training Accuracy	0.9970	0.9992	0.9782	0.75

The model validation accuracy reached 95% in BCNN better than traditional CNN when the epoch was set at 1000 whereas traditional convolutional neural network (CNN) reported 87.58% accuracy which is less in comparison, this may be due to gradient decent and ELBO that captured underfitting. BCNN reported less validation loss compare to CNN, this implies that BCNN can be recommended for androids apps for images detection and recognition in computer vision. The efficiency of BCNN over CNN may be due to ability of BCNN in capturing few samples of images. The study observed that BCNN does not overfit since the accuracy of its validation set is closed to the training set while traditional CNN was overfitted since the accuracy of its validation set is less than the training set. This affirms the superiority of BCNN over CNN. At 100 epoch BCNN reported 94% and 1.74 whereas CNN reported 84% and 0.457 validation accuracy and loss respectively.

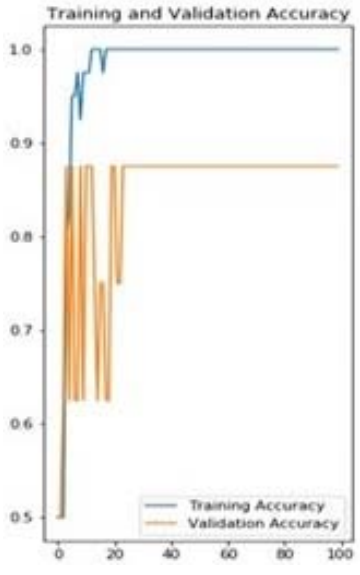

Fig. 3 Accuracy of Training and Validation sets of CNN

Table II depicts characteristics of the model parameter both in input and output.

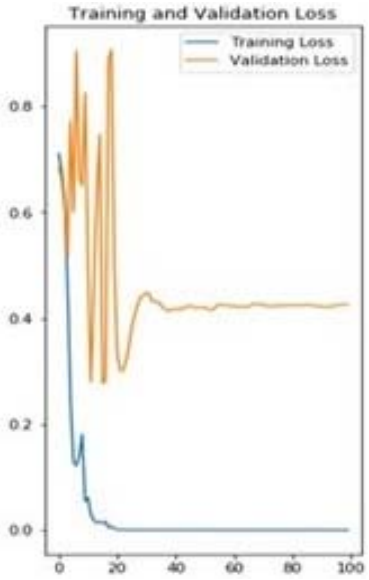

Fig. 4 Loss of Training and Validation sets of CNN

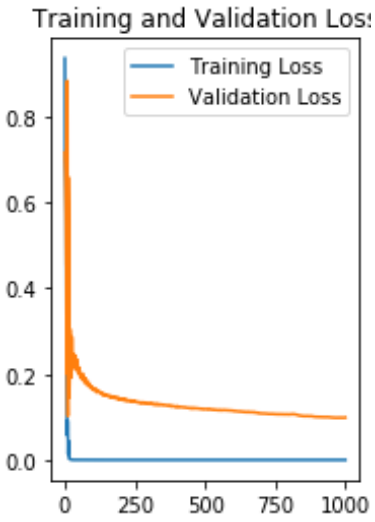

Fig. 5 Loss of Training and Validation sets of BCNN

TABLE II
CHARACTERISTICS OF NEURAL NETWORKS

Layer(type)	Output Shape	Parameters#
Conv2d(conv2d)	(None, 50, 50, 16)	448
max_pooling2d (MaxPooling2D)	(None, 25, 25, 16)	0
conv2d_1 (Conv2D)	(None, 25, 25, 32)	4640
max_pooling2d_1 (MaxPooling2)	(None, 12, 12, 32)	0
conv2d_2 (Conv2D)	(None, 12, 12, 64)	18496
max_pooling2d_2 (MaxPooling2)	(None, 6, 6, 64)	0
flatten (Flatten)	(None, 2304)	0
dense (Dense)	(None, 512)	1180160
Dense_1(Dense)	(None, 1)	513
Total Param:	1,204,257	
Trainable params:	1,204,257	
Non- Trainable params:	0	

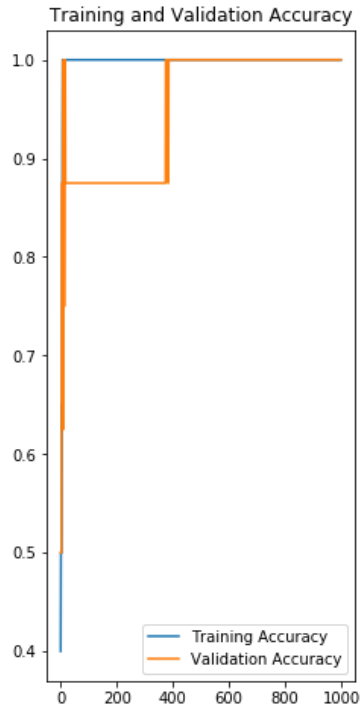

Fig. 6 Accuracy of Training and Validation sets of BCNN

IV. CONCLUSION

The study investigates the accuracy and loss of deep learning algorithms with the set of coronavirus (COVID-19) images dataset by comparing BCNN and traditional CNN in low dimensional dataset. The study found out that Bayesian convolution neural network outperformed conventional neural network at low dimensional dataset that could have exhibited under fitting.

REFERENCES

[1] A. Martín, A. Ashish, B. Paul, B. Eugene, C. Chifeng, C. Craig, S.C. Greg, D. Andy, D. Jeffrey, D. Matthieu, G. Sanjay, G. Lan, H. Andrew, L. Geoffrey, L. Michael, J. Rafal, J. Yangqing, K. Lukasz, K. Manjunath, L. Josh, M. Dan, S. Mike, M. Rajat, M. Sherry, M. Derek, O. Chris, S. Jonathon, S. Benoit, S. Ilya, T. Kunal, T. Paul, V. Vincent, V. Vijay, V. Fernanda, V. Oriol, W. Pete, W. Martin, Y. Yuan, and Z. Xiaoqiang, 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org, 2015.

[2] G. Yarin, I. Riashat and G. Zoubin, 'Deep Bayesian Active Learning with Image Data', Workshop on Bayesian Deep Learning, Neural Information Processing Systems, Barcelona, Spain, 2016.

[3] J. Schmidhuber, 'Deep learning in neural networks: "An overview', Neural Networks, Vol 61, pp 85-117, 2015.

[4] Y. Li. and Y. Liang, 'Learning over parameterized neural networks via stochastic gradient descent on structured data,' inadvances in Neural Information Processing Systems (NeurIPS), 2018.

[5] A. Kendall and Y. Gal, 'What uncertainties Do We need in Bayesian deep learning for computer vision?,' in Advances in Neural Information Processing Systems (NIPS), 2017.

[6] H. Jonathan and K. Nal, 'Bayesian Inference for Large Scale Image Classification', arXiv:1908.03491v1 (cs.LG), 2019.

[7] D. Giacomo, B. Christopher and Z. Xian, 'Bayesian Neural Networks for Cellular Image Classification and Uncertainty Analysis', bioRxiv preprint doi: https://doi.org/10.1101/824862, 2020.

[8] T. Toan, P. Trung, C. Gustavo, P. Lyle and R. Lan, 'A Bayesian Data Augmentation Approach for Learning Deep Models', 31st Conference

on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[9] F.K. Chen, R. D. Viljoen, and D.M. Bukowska,'Classification of image artefacts in optical coherence tomography angiography of the choroid in macular diseases'. Clinical & Experimental Ophthalmology, 44(5), 388–399. doi:10.1111/ceo.12683 2015.

[10] M. Krasser, 'Variational inference in Bayesian neural networks'. http://krasserm.github.io/2019/03/14/bayesian-neural-networks/ unpublished, 2019.

[11] M. Takashi, 'Bayesian deep learning: A model-based interpretable approach nonlinear Theory and Its Applications', IEICE, vol. 11, no. 1, pp. 16–35 c_ IEICE 2020 DOI: 10.1587/nolta.11.16

[12] R. G. Van and F. L. Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam. 1995

[13] Anaconda Software Distribution. Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/2020.