A Deep Learning Approach to Classifying Intracranial Hemorrhages

Emily Anaya[*]
Department of Electrical Engineering
Stanford University
eanaya@stanford.edu

Michael Beckinghausen[*]
Department of Chemical Engineering
Stanford University
mbecking@stanford.edu

Abstract

Urgent diagnosis of hemorrhage type and subsequent treatment is necessary for improved chances of survival for patients with brain hemorrhages. Machine learning models have been shown to be highly capable of assisting clinicians with the classification of intracranial hemorrhages. In this paper, we evaluate several 2-dimensional (2D) convolutional neural networks (CNNs) to perform multilabel 6-class classification (normal, epidural, intraparenchymal, intraventricular, subarachnoid, and subdural). We analyze our results using precision, recall, and f1-score evaluation metrics. Our most successful model was Mobile Net and we were capable of achieving an accuracy of 76% and a recall to determine existence of a hemorrhage of 93%.

1 Introduction

The problem that we have investigated is the detection and classification of intracranial hemorrhages. Diagnosing intracranial hemorrhages is an important challenge in the medical field as they can be fatal. Intracranial hemorrhage is bleeding that occurs inside the cranium. These hemorrhages account for approximately 10% of strokes in the U.S., the fifth-leading cause of death. Identifying the location and type of any hemorrhage present is a critical step in treating the patient. Traditional classification methods involve visual inspection by radiologists and quantitative estimation. The process is time-consuming and requires highly trained radiologists. A robust and efficient automated hemorrhage detection and classification algorithm is extremely valuable to clinics. An algorithm, such as the one we describe, that is capable of determining the type of hemorrhage would help localize the hemorrhage and radiologists' efficiency could significantly improve.

The current method of diagnosing intracranial hemorrhages involves taking a computed tomography (CT) scan of the brain, which is then analyzed by doctors. CT scans produce high contrast images due to the difference in x-ray absorption properties of brain tissue, blood, muscle, and bone [1]. The different x-ray attenuation values shown in CT images are expressed in Hounsfield units. CT images can be input into machine learning models such as CNNs to aid radiologists in analyzing them for hemorrhages.

CNNs have proven to be very successful in image classification tasks due to their ability to learn high-level image features automatically [1]. This has caused CNNs to become the leading machine learning architecture in image recognition tasks. We take advantage of CNNs to classify images containing hemorrhages. Using deep learning methods may assist radiologists in detecting subtle hemorrhages that can be difficult for radiologists to identify on their own.

[*] Equal Contribution

The rest of this paper is organized as follows. In Section 2, we present related works that have been implemented using convolutional neural networks. Feautures of our dataset are presented in Section 3. Section 4 describes our methods and Section 5 describes our results. Finally, we conclude our work with final thoughts and plans for future work in Section 6.

2 Related work

Several groups have experimented with various CNN architectures in order to classify intracranial hemorrhages. Previous groups have used either a 2-dimensional (2D) CNNs [2, 3, 4] or 3-dimensional (3D) CNNs [5, 1] to perform hemorrhage detection tasks. One group used the 2D approach, performing transfer learning with GoogLeNet and Inception-ResNet [3]. They trained only the last fully connected layers to achieve an accuracy of 0.982 and 0.992 respectively [3]. Another group performed a binary classification of hemorrhages using a 3D model called RADnet [5]. Their architecture incorporates a bidirectional LSTM layer to DenseNet-A [5]. This method declared the 3D CT scan as positive for hemorrhages if the model predicted a hemorrhage in three or more consecutive slices. RADnet achieved 81.82% accuracy, 88.64% recall, 81.25% precision, and 84.78% F1 score [5]. There is potential that using 3D blocks of stacked images as inputs rather than 2D images could lead to more accurate results as there is a loss of spatial information when the brain volume is analyzed using 2D slices rather than 3D [1].

Despite 3D CNNs potentially having a benefit over 2D CNNs, there are clear advantages in using a 2D CNN model, which we have implemented in this report. For example, 3D CNNs are more computationally intensive to train than 2D CNNs. Less computational cost and time is a significant benefit for clinical translation and applications [1].

Most previous works simply detect the presence or absence of a hemorrhage and do not classify the different types of hemorrhages. One group performed both detection and classification by training a 3D CNN to classify CT brain scans into normal scans (N) and abnormal scans containing subarachnoid hemorrhage (SAH), intraparenchymal hemorrhage (IPH), acute subdural hemorrhage (ASDH) and brain polytrauma hemorrhage (BPH) [1]. This groups' results are impressive, achieving F1-Scores of Normal: 0.819, SAH: 0.639, IPH: 0.427, ASDH: 0.829.

Image thresholding is commonly used prior to inputting the images to the machine learning model in order to optimize the detection of hemorrhages on CT brain scans [1, 5, 6]. One group showed a great improvement in F1 score for detecting hemorrhages when thresholding was applied (F1 score range improved from 0.706 - 0.902 to 0.919 - 0.952) [1]. Applying thresholding appears to significantly improve results. This is a technique that could improve the results of the method described in this report.

Additional state-of-the-art examples include a method using a pretrained AlexNet model [4], which achieves a hemorrhage recognition rate of 92.13%. Another group obtained a high test accuracy of 97% using a custom hybrid 3D/2D mask region of interst (ROI)-based convolutional neural network architecture for hemorrhage evaluation [6].

3 Dataset and Features

The input to our algorithm consists of CT images in dicom format where one image can contain at most three different types of hemorrhages and corresponding labels that specifiy the types of hemorrhages present. This labelled data set was obtained from Kaggle's RSNA Intracranial Hemorrhage Detection competition [7]. We used a CNN to output the predicted hemorrhage type, if one exists in the image. There are five types of intracranial hemorrhages that exist in the dataset: intraparenchymal, intraventricular, subarachnoid, subdural, and epidural. These types correspond to different bleeding locations. Images of the hemorrhage types can be seen in Figure 1, adapted from a figure located on Kaggle's RSNA Intracranial Hemorrhage Detection competition website [7].

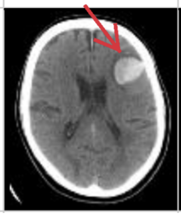
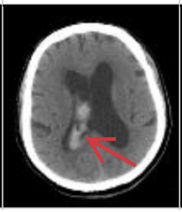
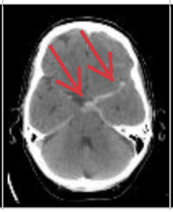
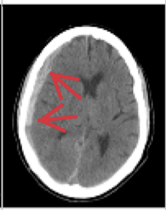
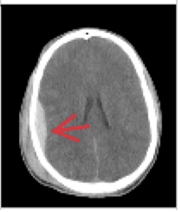
	Intraparenchymal	**Intraventricular**	**Subarachnoid**	**Subdural**	**Epidural**
Location	Inside of the brain	Inside of the ventricle	Between the arachnoid and the pia mater	Between the Dura and the arachnoid	Between the dura and the skull
Imaging					

Figure 1: Figure depicting the five types of hemorrhages that exist in the dataset and where they are located.

As can be seen in Figure 1, certain hemorrhages are very subtle to identify in the CT images. Epidural and subdural hemorrhages are especially difficult to see due to the low contrast that results from being adjacent to the bone of the head, which is of similar intensity to the hemorrhages. This challenge could cause these particular hemorrhages to be more difficult to identify by the algorithm.

Despite having access to a total of 674,262 images, we were limited to using only 5,000 images for training and testing due to a memory limitation. Our access to AWS training nodes failed to work and loading our dataset onto AWS resulted in a memory error. After meeting with TAs multiple times, we still could not resolve the issue. This caused us to train all of our models using a Desktop with a GeForce GTX 1060 6GB GPU, which was capable of training smaller models on a reasonable time scale. Ultimately, this resulted in us struggling with overfitting issues, since we could not easily address it by adding more data.

We conducted data preprocessing by manipulating the dicom format of the original images stored as (512,512) dicom arrays into png images of size (224,224). Additionally, all of the images were converted to 3-channel images. This was done so that transfer learning using MobileNet and other CNNs that require 3-channel images would be relatively easy, overcoming the need to change how the CNN models accept input images.

We experimented with two different dataset distributions. One dataset distribution was a 50-50 split of images containing and not containing hemorrhages. Another dataset distribution we experimented with disregarded images considered normal and was composed solely of images containing at least one hemorrhage. This essentially changed our problem to a multilabel 5-class classification on the hemorrhage types. This was done to improve our capability of determining the hemorrhage(s) types. Each dataset was respectively split into training and testing set with a ratio of 9:1. The training set was further split into training and validation with an 8:2 ratio.

4 Methods

We tested several models using transfer learning as well as created our own model using Keras. We used transfer learning to import the weights and structure of MobileNet and ResNet50. We chose binary cross entropy for our loss function because it is most suitable for our dataset.

Binary Cross Entropy Loss Equation:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{1}$$

We froze all of the pre-existing layers and added 2 dense layers with 64 nodes, 2 dropout layers with a 30 percent dropout rate, and finally a dense output later with 6 nodes. We chose the number of frozen layers and certain hyper-parameters such as a batch size of 32 and 60 epochs based on literature review of previous works [1, 3, 5], but we found that the model would usual fit the training data before 45 epochs.

5 Experiments/Results/Discussion

After trying a simple logistic regression model to perform classification and obtaining predictions entirely consisting of zeros, we decided to try a deeper network. We used this deep network as our baseline model. Specifically, the model is of the form: LINEAR -> RELU -> LINEAR -> SIGMOID. This method performed moderately obtaining 76 percent accuracy on the 50/50 dataset and 71 percent on the hemorrhage only dataset.

In a search to achieve higher prediction accuracy, we compared performances of several CNNs using transfer learning, including ResNet50 and MobileNet, each trained using Adam and SGD optimizers. The residual blocks of ResNet50 contain skip connections, allowing the deep learning model to learn the identify function very easily. Therefore, in general, resnet blocks can be added without hurting the performance of the model. In addition, these skip connections reduce the risks of vanishing gradients which can occur in deep neural networks [8].

We noticed our model would obtain a high accuracy on the training set, but when we re-ran, the training set would have a much lower accuracy. After quite a bit of research we were found an article on the issue of using batch normalization on transfer learning in Keras [9]. There is a current issue where if a batch norm layer is left not trainable it will train as if it is updating, but when it is used to predict, the layer uses its original values from the imported model. The solution to this is to make all of the batch normalization layers trainable. While this fixed one issue, it also presented a new problem. The model now would overfit due to the large number of batch normalization layers in ResNet and MobileNet. This issue more strongly affected ResNet, since it has more batch normalization layers, and resulted in very inaccurate test predictions. While we conducted all of the same tests on ResNet, we have opted to leave it out of the figures in the report.

We decided to modify the number of layers that we appended to the model. We implemented dropout and reduced the size of network to help with the issue, but found that it had little to no effect on the overfitting we experienced. After rigorously trying various architecture designs and sizes and changing the parameters of dropout we found that the issue of overfitting still remained. Thus, we decided to create our own sequential model and train that. We based the design off of [10]. Our results are shown in table 1, comparing the F1, precision, and recall performance of our baseline model, ResNet, and MobileNet models.

We wanted to use a deep enough neural network that was capable of learning features necessary for the complex task of hemorrhage classification. However, our limiting factors in increasing model complexity was computational time and memory.

The performance metric that we were most interested in maximizing is recall. This metric is particularly important in hemorrhage classification because the consequences of missing a brain hemorrhage can be severe and deadly. The cost of lower precision is accepted for medical applications that have follow-up diagnosis. We are also interested in the precision and F1-Score, a metric that takes both precision and recall into account. Below, we show the equations that define these metrics.

Precision:

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall:

$$R = \frac{TP}{TP + FN} \tag{3}$$

F1-Score:

$$F1 = 2\frac{PxR}{P + R} \tag{4}$$

Our precision, recall, and F1 scores for each model using the 50-50 dataset is shown in Table 1. Our results using the dataset containing solely hemorrhage images is shown in Table 2.

The results show that epidural hemorrhages were the most difficult hemorrhages to detect. We believe this is because this particular hemorrhage is located adjacent to the bone, which is of similar intensity to the hemorrhages. Some interesting results to note are that no one model completely outperforms another. For example MobileNet has the best recall for detecting any hemorrhage, but it is much worse than Our Model at predicting intraventricular hemorrhages.

Table 1: Model Performance Comparison Using a 50-50 Split Dataset

	Baseline			Our Model			MobileNet		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Epidural	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Intraparenchymal	0.04	0.33	0.03	0.24	0.16	0.44	0.23	0.17	0.36
Intraventricular	0.05	0.11	0.03	0.20	0.14	0.38	0.08	0.06	0.15
Subarachnoid	0.02	0.24	0.22	0.27	0.18	0.57	0.25	0.18	0.44
Subdural	0.27	0.24	0.32	0.21	0.14	0.45	0.27	0.21	0.41
Any	0.46	0.46	0.46	0.61	0.53	0.73	0.65	0.50	0.93

Table 2: Model Performance Comparison Using a Solely Hemorrhage Containing Dataset

	Baseline			Our Model			MobileNet		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Epidural	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Intraparenchymal	0.30	0.28	0.33	0.26	0.21	0.35	0.41	0.29	0.70
Intraventricular	0.30	0.33	0.35	0.40	0.31	0.57	0.11	0.12	0.10
Subarachnoid	0.38	0.29	0.52	0.45	0.33	0.71	0.44	0.30	0.83
Subdural	0.42	0.60	0.32	0.60	0.47	0.80	0.64	0.49	0.91
Any	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

6 Conclusion/Future Work

In this paper, we present several convolutional neural network models to achieve high classification performance of intracranial hemorrhages. MobileNet performed the best with an impressive recall of 0.91 for subdural hemorrhages and 0.83 for subarachnoid. We believe that this algorithm performed better than our model because we were able to utilize transfer learning and its additional complexity made it more capable of learning our training set.

We plan to try implementing a 3D CNN model in our future work and compare the performance to the 2D CNN model. We believe a 3D CNN model could lead to improved results because it takes into account the depth information of the hemorrhages. Another interesting idea would be to include images in the axial, sagittal, and coronal planes of the human. Similar to the 3D CNN, we believe this would provide additional, useful information to our classification algorithm, leading to improved results.

We also plan to normalize the CT images based on mean and standard deviation before passing them into the model. This technique is useful when the images are obtained from different machines to address light disparities between the images. Additionally, we would like to apply a preprocessing threshold technique similar to the one described in [1].

Finally, and objectively most importantly, we plan to train the model on the full dataset and not just the 5000 images that we were restricted to when conducting our training. We expect that this will help to address overfitting and allow us to achieve much higher precision and recall.

7 Contributions

Both contributors implemented the code and assisted in writing the paper. Michael worked on code implementation and paper. Emily worked on performance metrics code and paper and poster. Most coding sessions where conducted together, and decisions on what modifications to make were always discussed before proceeding.

References

[1] Satya Singh. Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans. *Sensors*, 19:2167, 05 2019.

[2] A. Majumdar, L. Brattain, B. Telfer, C. Farris, and J. Scalera. Detecting intracranial hemorrhage with deep learning. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 583–587, July 2018.

[3] Tong Duc Phong, Hieu N. Duong, Hien T. Nguyen, Nguyen Thanh Trong, Vu H. Nguyen, Tran Van Hoa, and Vaclav Snasel. Brain hemorrhage diagnosis by using deep learning. In *Proceedings of the 2017 International Conference on Machine Learning and Soft Computing*, ICMLSC '17, pages 34–39, New York, NY, USA, 2017. ACM.

[4] Muhammad Awwal Dawud, Kamil Yurtkan, and Huseyin Oztoprak. Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. *Computational Intelligence and Neuroscience*, 2019:1–12, 06 2019.

[5] Monika Grewal, Muktabh Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. pages 281–284, 04 2018.

[6] P.D. Chang, E. Kuoy, J. Grinband, Brent Weinberg, M. Thompson, Richelle Homo, J. Chen, H. Abcede, Marzie Shafie, L. Sugrue, C.G. Filippi, M-Y Su, W. Yu, C. Hess, and D. Chow. Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. *American Journal of Neuroradiology*, 39, 07 2018.

[7] Rsna intracranial hemorrhage detection.

[8] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2016.

[9] Vasillis Vryniotis. The batch normalization layer of keras is broken, 2018.

[10] Sarkar. A comprehensive hands-on guide to transfer learning with real-world applications in deep learning, Nov 2018.