Doubly Weak Supervision of Deep Learning Models for Head CT

Khaled Saab[1]*, Jared Dunnmon[2]*, Roger Goldman[3], Alex Ratner[2], Hersh Sagreiya[3], Christopher Ré[2], and Daniel Rubin[4]

[1] Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[2] Department of Computer Science, Stanford University, Stanford, CA, USA
[3] Department of Radiology, Stanford University, Stanford, CA, USA
[4] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
{ksaab,jdunnmon,goldmanr,ajratner,sagreiya,chrismre,dlrubin}@stanford.edu

Abstract. Recent deep learning models for intracranial hemorrhage (ICH) detection on computed tomography of the head have relied upon large datasets hand-labeled at either the full-scan level or at the individual slice-level. Though these models have demonstrated favorable empirical performance, the hand-labeled datasets upon which they rely are time-consuming and expensive to create. Further, given limited time, modelers must currently make an explicit choice between scan-level supervision, which leverages large numbers of patients, and slice-level supervision, which yields clinically insightful output in the axial and in-plane dimensions. In this work, we propose doubly weak supervision, where we (1) weakly label at the scan-level to scalably incorporate data from large populations and (2) model the problem using an attention-based multiple-instance learning approach that can provide useful signal at both axial and in-plane granularities, even with scan-level supervision. Models trained using this doubly weak supervision approach yield an average ROC-AUC score of 0.91, which is competitive with those of models trained using large, hand-labeled datasets, while requiring less than 10 hours of clinician labeling time. Further, our models place large attention weights on the same slices used by the clinician to arrive at the ICH classification, and occlusion maps indicate heavy influence from clinically salient in-plane regions.

Keywords: Weak Supervision · Multiple Instance Learning · Head CT

1 Introduction

Non-contrast computed tomography of the head (HCT) is the most commonly used first-line diagnostic imaging method for patients presenting with acute neurologic deficits or head trauma [3]. The problem of rapidly detecting such pathologies as intracranial hemorrhage (ICH) on HCT represents an important task in clinical radiology to expedite clinical triage and care. For example, in

* Equal contribution.

the case of a stroke, doctors must quickly rule out hemorrhage before beginning treatment with tissue plasminogen activator to break down clots. Encouragingly, machine learning algorithms based on deep convolutional neural networks (CNNs) have recently demonstrated high levels of performance on automated ICH detection [1, 2, 9, 10]. However, development of these models has relied on large datasets hand-labeled at either the scan-level or the individual slice-level, each of which requires significant investment of domain expert labeling time. For perspective, Chang et al. [1] use 512,598 slices from 10,159 scans annotated with voxel-level segmentation masks by a board-certified radiologist, Lee et al. [10] use 14,758 slices from 904 scans hand-labeled at the slice-level by five separate board-certified neuroradiologists, Chilamkurthy et al. [2] use subsets of a total of 313,318 scans hand-labeled at the individual slice-level, and Jnawali et al. [9] use 40,367 scans labeled at the scan-level.

The different supervision strategies pursued in existing work point to an underlying trade-off that has important implications in a clinical environment. While expending a fixed budget of labeling effort at the scan-level allows models to leverage larger patient populations that could improve generalization, supervising at the slice-level yields models that output useful diagnostic information in both axial and in-plane dimensions [1, 2, 9, 10]. Slice-level and voxel-level information could be particularly useful to clinicians for purposes ranging from slice prioritization while reading to assessment of the clinical relevance of model predictions. In this work, we propose a method by which one can combine two separate techniques for using weaker, or noisier, supervision – generative modeling of domain-specific heuristics to create scan-level labels, and discriminative modeling within a multiple-instance learning (MIL) framework – to rapidly train deep learning models for ICH detection. We present evidence that this approach can provide the best of both worlds: supervision at the scan-level that is cheap to acquire for large populations, along with axial and in-plane output that can improve both practical utility and interpretability.

With a total investment of under 10 hours of clinician time and a modestly sized dataset, our approach achieves an average Area Under the Receiver Operating Characteristic Curve (ROC-AUC) score of 0.91 on ICH detection. This result is comparable to existing results using large hand-labeled datasets. Moreover, our model formulation enables insightful information to be extracted at a level more granular than the provided supervision, as we can use the attention weights themselves to identify important axial slices and standard network occlusion techniques to quickly identify salient in-plane regions. Our results point to a promising approach for developing clinically useful classification models that require minimal clinician labeling time.

2 Related Work

Deep learning models such as CNNs are capable of extracting meaningful patterns from raw inputs for a wide variety of medical imaging tasks, ranging from cancer diagnosis [5] to chest radiograph worklist prioritization [4]. Several re-

Table 1. Summary of recent studies on ICH detection with deep learning [1, 2, 9, 10].

Source	Dataset Size	Labeling	Model	Explainability	ROC-AUC
Lee et al.	14,758 slices	Slice	2D CNN	Heatmaps	0.96
Chang et al.	512,598 slices	Slice	3D/2D CNN	Bounding boxes	0.98
Chilamkurthy et al.	165,809 slices	Slice	2D CNN	Segmentations	0.93
Jnawali et al.	40,357 scans	Scan	3D CNN	None	0.86
Ours	4,340 scans	Scan	2D CNN	Occlusion Maps	0.91

searchers have recently utilized large, hand-labeled datasets labeled at varying degrees of granularity to achieve high levels of performance for ICH detection on HCT; Table 1 summarizes these results in the context of the present work.

We propose using weak supervision methods to rapidly train ICH detection models, as these techniques can substantially reduce required labeling time. Weak supervision formally refers to using weaker, or noisier signals to supervise training of machine learning models. Notable work that uses weak supervision approaches for medical imaging includes that of Wang et al. [13] on chest radiograph classification and that of Fries et al. [6] on rare disease detection from cardiac magnetic resonance imaging (MRI). In this study, we leverage a type of weak supervision referred to as data programming [11], in which a generative modeling technique is used to combine user-defined heuristics written over unlabeled data to create probabilistic training labels for deep neural networks.

From a discriminative modeling standpoint, we leverage the idea of MIL in this work. MIL refers to the setting wherein an example consisting of a collection of instances is considered positive if and only if at least one instance in that collection is positive. Some successful implementations of MIL in medical imaging include histopathology segmentation [14] and myocardial infarction detection [12]. We specifically utilize the attention-based MIL technique of Ilse et al. [8], where an attention layer is trained to aggregate multiple instances (e.g. CT slices) in order to give an overall bag-level classification (e.g. CT scan).

3 Dataset and Preprocessing

A collection of 5,582 non-contrast HCT studies, which were performed over the timespan of 2001 through 2014, was acquired from our institutional PACS under an approved IRB. We considered studies containing more than 29 and less than 44 axial CT slices with 5mm axial resolution. Each series with less than 44 slices was padded with additional homogenous images with Hounsfield Unit values of 0 such that they contained 44 CT slices. The center 32 slices were then selected for automated analysis. The resulting dataset consisted of 4,340 CT studies with accompanying text reports that could be used for weak label extraction via data programming, as described in Methods. We randomly selected 4,000 of these as a training set, and obtained scan-level hand-labels for the remaining 340 determined through the consensus of two radiology fellows. These 340 hand-labeled exams were evenly split between development and test sets, and included

44 positive examples for ICH. A single radiologist provided segmentations for 15 of these ICH cases to enable analysis at the in-plane level.

4 Methods

In this work we aim to demonstrate that with minimal required clinician input (compared to the curation of large hand-labeled datasets), we are able to develop clinically useful classification models through a doubly weak supervision process, which we describe in detail below.

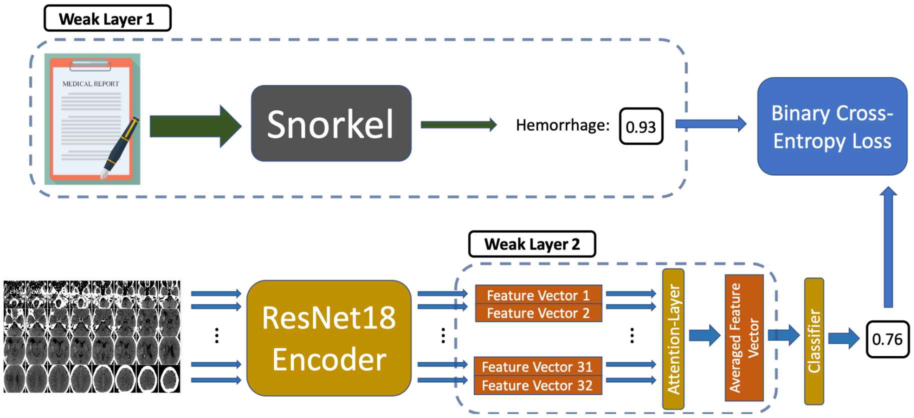

Fig. 1. In our doubly weak supervision approach for hemorrhage detection on head CT, we first use data programming to extract weak scan-level labels from clinical text reports. We then use a shared ResNet-18 image encoder to compute a feature vector for each CT slice. This series of feature vectors is aggregated using an attention layer. The aggregated feature vector is then fed into a classifier, and the scan-level model output is compared with the weak label from the text via a binary cross-entropy loss.

4.1 First Weak Supervision Layer: Data Programming

To reduce the amount of labeling time required to train our ICH detection model, we use a method of weak supervision known as data programming, wherein users write heuristic labelling functions (LFs) that programmatically label training data [11]. In our case, a single radiology fellow wrote seven LFs over the text reports to support generation of weak labels. The LFs consisted of regular expressions that look for phrases in text that either indicate the existence or absence of ICH. For example, one LF outputs a positive label if the text includes the words “hemorrhage” or “hematoma,” but does not have nearby negating words like “no” or “without.” In contrast to the costly approach of manually annotating thousands of scans, this process took under 10 hours of clinician development time using a hand-labeled development set of 350 reports for LF tuning. Probabilistic labels were then automatically generated from these LFs for the remaining unlabeled examples using statistical modeling tools in the

Snorkel data programming software package, the theory and implementation of which are described in Ratner et al. [11].

Labels generated for the 350-report validation set using data programming achieved an F1 score of 0.95 with respect to human annotation. These results indicate that while noise does exist in our generated labels, they are generally of high quality.

4.2 Second Weak Supervision Layer: Multiple-Instance Learning

In addition to obtaining weak scan-level labels from the text, we use these coarse scan-level labels to supervise an MIL model that can naturally produce slice and voxel-level outputs. HCT classification fits naturally within the MIL paradigm because we have a single label per CT scan, which comprises a collection of 32 image slices. Further, if a CT scan is labeled as positive for ICH, then there must exist at least one slice that contains evidence of ICH. We therefore adopt an attention-based deep MIL algorithm proposed by Ilse. et al. [8], which also leads to natural interpretability in the axial dimension via attention weights computed for each axial slice. From an implementation standpoint, we first embed each CT slice into a feature space of dimension 50 using a ResNet-18 encoder [7]. This ResNet-18 architecture was chosen because it is a standard CNN that performs well on computer vision tasks, and generalizes just as well to our development set as deeper variants such as ResNet-50 or ResNet-101. Second, we pass each feature vector into an attention layer (a two-layer fully connected neural network), which outputs a weight for each feature vector. These weights are used to compute a weighted average of the 32 feature vectors for each CT scan. The resulting averaged feature vector is then passed through a classifier, which in our case is a single fully-connected layer followed by a sigmoid activation function (Fig. 1).

4.3 Model Training Procedure

Each slice is downsampled to dimensions of 224 x 224 and is normalized using a global mean and standard deviation. We then feed each slice to our ResNet-18 encoder, which is randomly initialized. We train our models using stochastic gradient descent with an initial learning rate of 0.1, which is reduced by a factor of 10 at each epoch if a plateau in the validation loss is observed. The weight decay parameter, which was found using a coarse hyperparameter search, is set to 0.005, and we use the largest batch size that fit into our Tesla P100 GPU, which was 12 HCT scans. All other parameters are set to their default values. Model training is performed using the PyTorch 0.4.1 software library, and each model took approximately six hours to train over 50 epochs.

Due to the low number of positive cases in the ground-truth dataset, model performance on the test set could be sensitive to the random splitting of the development and test set. We therefore carry out a five-fold cross-validation procedure where we repeat a random stratified splitting of the 340 hand-labeled scans into development and test sets, and report average model performance over five trials with different random seeds.

5 Results and Discussion

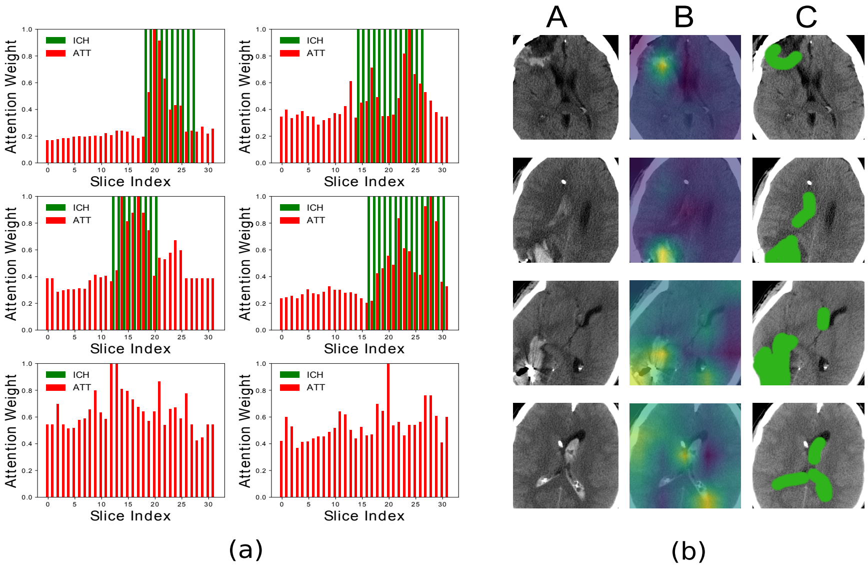

Fig. 2. (a) Normalized attention weights ("ATT," red) compared to ground-truth slice annotations for ICH ("ICH," green); the bottom two cases have no ICH. (b) ICH appears in raw CT slices (column A) as a hyperdense (i.e. whiter) substance as compared within the intracranial vault to the surrounding parenchyma and fluid. Voxel-level hand labels (column C) coarsely identify voxels most likely to represent ICH. Occlusion maps (column B) demonstrate favorable correlation with the coarsely labeled voxels.

5.1 Evaluating Overall ICH Detection Performance

The first fundamental hypothesis underlying this work is that weakly supervised MIL can yield high levels of performance on automated detection tasks on HCT. We evaluate if this is the case by computing the mean ROC-AUC value of our ICH classifier and assessing if this number is competitive with those presented in the literature. As shown in Table 1, the 0.91 average ROC-AUC value, with standard deviation 0.017, obtained for ICH detection from the weakly supervised CNN-MIL model is comparable to existing results in the literature, even when using a modestly sized dataset. Importantly, we obtained these results using only 10 hours of clinician time spent on labeling, and our labeling approach is fundamentally scalable; to improve performance, we could increase the size of the dataset by obtaining additional unlabeled data (we did not have more data available for this study), evaluating our data programming model over the text reports, and training our CNN-MIL model on the resultant probabilistic labels.

5.2 Analyzing Axial Attention Weights

Our second hypothesis is that even though we supervise at the scan-level, we are still able to extract meaningful slice-level information from our trained model. At the core of our attention-based MIL model is a 2-layered neural network (i.e. the attention layer) that outputs 32 numbers, which are used to take a weighted average of all 32 feature vectors representing the CT slices. The output of this attention layer therefore represents how much weight the model places on each CT slice to make the overall scan-level classification. To assess the behavior of these axial attention weights, we compute their Gini coefficient, which measures the inequality amongst the attention weight values. We find a strong positive correlation between the Gini coefficient and the model prediction, with a Pearson correlation coefficient of 0.97 (see Supplementary Material). This indicates that the lower the probability of hemorrhage (i.e. less likely the model predicts existence of ICH), the more uniform the attention weights will be (i.e. lower Gini coefficient), which is what we expect since a negative case would not have a hemorrhage on any axial slice.

Additionally, we qualitatively analyze these learned weights by comparing them to the ground truth slice-level annotations of four random true positive cases and two random true negative cases. As shown in Fig. 2(a), across the four random true positive cases, we observe that the largest attention weights (in red) in each case do indeed occur within the band of slices used by the clinician to make the ICH classification (in green). Quantitatively, the accuracy of the top 20% of slices with the highest attention weights with respect to the slices with positive ground-truth segmentations[1] was 83%, while the accuracy of the top 10% of slices was 93%. Moreover, while we did not constrain the model to output attention weights that are similar for spatially adjacent slices, those shown in Fig. 2(a) demonstrate smoothness in the axial direction, indicating that model predictions appropriately reflect the spatial localization of ICH. Such a result gives additional confidence that the CNN encoder is learning a representation that reflects clinically relevant features.

5.3 Interpreting In-Plane Signal

Our third hypothesis is that we can extract clinically meaningful in-plane information from our trained CNN-MIL model, even though it was weakly supervised at the study-level. To assess this outcome, we generate occlusion maps for each slice that indicate which regions most influence model output, and qualitatively evaluate the degree of overlap between these occlusion maps and the hand-labeled, voxel-level segmentation. Our occlusion maps are computed by iteratively obscuring 20x20 patches from each CT slice in a scan while recording the change in the hemorrhage probability emitted by the model. In Fig. 2(b), we show four random true-positive examples where we clearly see alignment between extracted occlusion maps and ground-truth segmentation maps. As shown

[1] Evaluated over the 15 cases with radiologist-provided segmentation.

in Table 1, this type of in-plane information has not been successfully extracted from ICH detection models supervised at the scan-level, and has been onerous to achieve for those supervised at the slice-level. For instance, while these occlusion maps are conceptually similar to the prediction report introduced by Lee et al. [10], our model did not require tens of thousands of expert-labeled slices.

6 Conclusion

In this work, we have used an attention-based MIL model weakly supervised at the scan-level to achieve ROC-AUC scores comparable with those of state-of-the-art methods, while using under 10 hours of clinician time. Further, we have shown that these models output meaningful information in axial and in-plane dimensions, which can have substantial clinical utility. These results point towards a promising approach to training high-performance volumetric imaging models that leverage large patient populations, provide useful output at high levels of spatial granularity, and make efficient use of clinician labeling resources.

References

1. Chang, P., Kuoy, E., Grinband, J., Weinberg, B., Thompson, M., Homo, R., Chen, J., Abcede, H., Shafie, M., Sugrue, L., et al.: Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. American Journal of Neuroradiology **39**(9), 1609–1616 (2018)
2. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018)
3. Coles, J.: Imaging after brain injury. British Journal of Anaesthesia **99**(1), 49–60 (2007)
4. Dunnmon, J.A., Yi, D., Langlotz, C.P., Ré, C., Rubin, D.L., Lungren, M.P.: Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology p. 181422 (2018)
5. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
6. Fries, J.A., Varma, P., Chen, V.S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., Delp, S., Ashley, E., Ré, C., Priest, J.R.: Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. Nature Communications **10**(1), 3111 (2019)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2132–2141 (2018)
9. Jnawali, K., Arbabshirani, M.R., Rao, N., Patel, A.A.: Deep 3D convolution neural network for CT brain hemorrhage classification. In: Medical Imaging 2018: Computer-Aided Diagnosis. vol. 10575, p. 105751C. International Society for Optics and Photonics (2018)

10. Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S., E. Guerrier, C., A. Ebert, S., Pomerantz, S., Romero, J., Kamalian, S., G. Gonzalez, R., Lev, M., Do, S.: An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nature Biomedical Engineering **3** (12 2018). https://doi.org/10.1038/s41551-018-0324-9
11. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Re, C.: Snorkel: Rapid training data creation with weak supervision. Proceedings VLDB Endowment **11**(3), 269–282 (Nov 2017)
12. Sun, L., Lu, Y., Yang, K., Li, S.: ECG analysis using multiple instance learning for myocardial infarction detection. IEEE Transactions on Biomedical Engineering **59**(12), 3348–3356 (2012)
13. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3462–3471 (2017)
14. Xu, Y., Zhang, J., Eric, I., Chang, C., Lai, M., Tu, Z.: Context-constrained multiple instance learning for histopathology image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 623–630. Springer (2012)