IMPROVED ICH CLASSIFICATION USING TASK-DEPENDENT LEARNING

Amir Bar¹, Michal Mauda², Yoni Turner³, Michal Safadi¹, Eldad Elnekave¹, ⁴

¹Zebra Medical Vision Inc.
²Tel-Aviv Medical Center, ³Shaare Zedek Medical Center, ⁴Rabin Medical Center

ABSTRACT

Head CT is one of the most commonly performed imaging studies in the Emergency Department setting and Intracranial hemorrhage (ICH) is among the most critical and time-sensitive findings to be detected on Head CT. We present BloodNet, a deep learning architecture designed for optimal triaging of Head CT’s, with the goal of decreasing the time from CT acquisition to accurate ICH detection. The BloodNet architecture incorporates dependency between the otherwise independent tasks of segmentation and classification, achieving improved classification results. AUCs of 0.9493 and 0.9566 are reported on held out positive-enriched and randomly sampled sets comprised of over 1400 studies acquired from over 10 different hospitals. These results are comparable to previously reported results with smaller number of tagged studies.

Index Terms—Deep Learning, Segmentation, ICH, Hemorrhage, Classification

1. INTRODUCTION

Intracranial hemorrhage (ICH) is a critical finding seen in various clinical circumstances spanning major trauma to spontaneous intracranial aneurysmal rupture. Early and accurate detection is essential in achieving optimal outcomes. An AI-facilitated first read of CT brains could provide value by detecting subtle bleeds which might go unrecognized, as well as providing triage-service to prioritize positively-flagged studies for expert radiologist review.

In recent years, convolutional neural networks (CNN’s) have been successfully designed to detect various pathologies in medical imaging [1] [2] [3] [4]. Previously reported deep-learning infrastructures for automatic ICH detection have based ICH prediction upon either the entire 3d Head CT volume [5] or each 2d CT slice [6] [7]. While the former potentially utilizes a larger amount of data, it is at the cost of relatively weak supervision due to the high dimensionality of the input volume. The second approach requires a substantial tagging effort due to tedious annotation of every relevant slice in the scan.

Jnawali et al [5] assembled a dataset of 40k studies and preprocessed it to a fixed input size. It was then used for the training of a 3D convolution [8] classification pipeline and reported to have an AUC of 0.86 using a single model. Additional work was in [6], in which the authors utilized a large dataset of 6k studies tagged slice-wise by radiologists for training. To localize the findings, the authors had to annotate the slices pixel-wise to create the masks necessary in order to train a UNet [9] architecture for segmentation. They report AUC of 0.9419 for the classification part. In [7], the authors used multiple segmentation auxiliary losses to leverage the pixel-wise information and aggregated the 3D volumetric decision using LSTM [10].

The present report describes integration of both classification and segmentation of an image in a single network, utilizing the pixel-wise prediction to improve the 3D volumetric ICH classification result. BloodNet is a CNN architecture which explicitly incorporates the pixel-wise prediction through modeling the dependency between the classification and segmentation task.

2. MATERIALS AND METHODS

For training and validation, 175 non-contrast CT brain studies with ICH-positive radiology reports were reviewed by at least one expert radiologist who validated the existence of the reported ICH and manually segmented it. An ICH-negative dataset including 102 CTs was also assembled. For validation we use only positive studies, which contain both positive
The present report describes a new pipeline for CT-based ICH classification intended for enhanced triage. The setup relies on the learning of both classification and segmentation, having demonstrated that the segmentation task provides synergistic support to the ICH classification task. A high level description of our architecture is described in Figure 5.

To exploit the volumetric nature of ICH, the input number of slices was set as 5 consecutive axial CT slices, allowing for better detection of true ICH. We empirically observed that the learned models better distinguish artifacts and hemorrhages, which may look similar on a single slice but commonly appear differently over consecutive slices. We show example for the advantage of such context in Fig 2. Additional preprocessing included the utilization of standard brain-windowing. Since we empirically observed that a hemorrhage might be very small, we kept the input slices in the full 512x512 CT resolution.

Given the input slices we first base our approach on performing classification alone, using the architecture in Figure 3. Hence our classification loss is:

$$ L_{\text{classification}} = \frac{1}{m} \sum_{i=1}^{m} CE(y_i, \hat{y}_i) $$

Where $y_i$ is the ground truth label, $\hat{y}_i$ is the prediction of the $i$-th sample, $m$ is the number of samples and $CE$ is the binary cross entropy function:

$$ CE(y, \hat{y}) = y \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}) $$

Considering the clear advantages of multi-task learning reported in recent research [11, 7], we modified the architecture and added a decoder to enable the multi-task learning scenario of classification and segmentation (see Figure 4). We also added an auxiliary segmentation loss:

$$ L_{\text{segmentation}} = \frac{1}{m \cdot h \cdot w} \sum_{i=1}^{m} \sum_{j=1}^{h} \sum_{k=1}^{w} CE(y_{ijk}, \hat{y}_{ijk}) $$

Where $h$ and $w$ are the height and width of input slice, $\hat{y}_{ijk}$ is the pixel in the spatial position $j, k$ of the $i$-th sample.

Our final loss is thus:

$$ L = (1 - \lambda) L_{\text{classification}} + \lambda \cdot L_{\text{segmentation}} $$

Finally, instead of implicitly using the segmentation information as supervision, we explicitly design the architecture to utilize the segmentation information to support classification. More specifically, we sum over the decoder network segmentation prediction, multiply by the voxel volume and concatenate the approximation of blood in $mm^3$ as a feature in the classification branch.

Table 1. Data

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>3953</td>
<td>1815</td>
</tr>
<tr>
<td>Negative</td>
<td>22122</td>
<td>4141</td>
</tr>
</tbody>
</table>

Every cell represents the number of tagged slices. All slices were manually pixel wise annotated for positive ICH on 175 ICH positive and 102 on ICH negative scans.

Fig. 2. Context is necessary with respect to ICH. Top row: Parenchymal hemorrhage. Bottom row: calcification.

Fig. 3. Single task, classification. Predicting whether an input has positive indication for ICH.

Fig. 4. Multi-task, classification and segmentation. The shared encoder learns a representation suitable for both tasks.
Fig. 5. Task dependent, classification is dependent on segmentation. The segmentation branch result probabilities are multiplied by the voxel volume of the input, summed and finally concatenated to the bottleneck as an additional feature for classification.

To train this architecture, we employ three steps. First, we train the segmentation branch alone. Then, we freeze all weights and train only the last fully connected layer of the classification branch. Finally, we train the entire architecture for both classification and segmentation in an end-to-end manner. Respectively, we use $\lambda = 1$, $\lambda = 0$, $\lambda = 0.5$ in the loss equation. In all our experiments we use the Adam optimizer with learning rate of $1e^{-4}$ and exponential decay of 0.96. In inference, given a study, we compute the probability for ICH over every slice and use the maximal probability as the study probability for ICH.

3. RESULTS

We choose the best architecture using AUC over validation set. Table 2 provides comparison between models. We then evaluated on two different held out test sets, a positive enriched and a randomly sampled sets. The advantage of a positive enriched set is in representation of different types of ICHs as well as ICHs which are less prevalent. To collect this set we used a textual search over radiology reports. Since such data collection method might present a bias towards a specific search criteria, we also collected a randomly sampled set. We assume that in the randomly sampled set the cases in radiologists daily routine are well represented. We report AUCs of 0.9493 and 0.9566 over the enriched and randomly sampled tests set. Table 2 provides further information. A manual review of false positives showed propensity to aberrantly misclassify calcified hemangiomas, dystrophic parenchymal calcifications and basal ganglial calcifications.

4. DISCUSSION

This work provides further evidence to support the approach of utilizing pixel wise annotated data for classification. However, our results indicate that relying on the multi-task setting alone might not be enough to yield a significant improvement in performance for classification. In BloodNet, we explicitly model a segmentation dependent classification, resulting in design that fully leverages the dense pixel wise supervision to boost classification performance. It has the advantage of both classification and localization of the acute finding and while classification is most important in a triage system, the local-
Table 3. Comparison between models

<table>
<thead>
<tr>
<th>Network</th>
<th>AUC</th>
<th>0.95 CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet50 [12]</td>
<td>0.9159</td>
<td>[0.9081, 0.9236]</td>
</tr>
<tr>
<td>ii. BloodNet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single task, classification</td>
<td>0.9453</td>
<td>[0.9395, 0.9512]</td>
</tr>
<tr>
<td>Multi task, classification and segmentation</td>
<td>0.9411</td>
<td>[0.9352, 0.9471]</td>
</tr>
<tr>
<td>Task dependent, segmentation dependent classification</td>
<td>0.9658</td>
<td>[0.9611, 0.9704]</td>
</tr>
</tbody>
</table>

Ablation studies of results over validation set slices. Mean AUC and CI are are computed using bootstrap ($n = 10^4$).

Fig. 6. Visualizations of studies which received high probability for ICH on validation set.

Fig. 7. Segmentation result example. While the first priority is to boost classification performance to support triage, accurate localization is byproduct of the proposed approach.

Fig. 8. Visualizations of negative studies which received high probability. Mengiomias and calcifications.

Acknowledgement

The authors would like to thank Orna Bregman, Assaf Pinhasi, Jonathan Laserson, David Chettrim, Chen Brestel, Eli Goz, Phil Teare, Tomer Meir, Rachel Wities, Amit Oved, Raouf Muhamedrahimov and Eyal Toledano for helpful comments and discussions during this research.
References


