applied sciences

MDPI

Article

Segmentation of Intracranial Hemorrhage Using Semi-Supervised Multi-Task Attention-Based U-Net

Justin L. Wang [1], Hassan Farooq [2], Hanqi Zhuang [3] and Ali K. Ibrahim [3,*]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA; jlwang5@illinois.edu
[2] ECE Department, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA; hfaroo9@illinois.edu
[3] CEECS Department, Florida Atlantic University, Boca Raton, FL 33431, USA; zhuang@fau.edu
* Correspondence: aibrahim2014@fau.edu

Received: 13 April 2020; Accepted: 6 May 2020; Published: 9 May 2020

Abstract: Intracranial Hemorrhage (ICH) has high rates of mortality, and risk factors associated with it are sometimes nearly impossible to avoid. Previous techniques to detect ICH using machine learning have shown some promise. However, due to a limited number of labeled medical images available, which often causes poor model accuracy in terms of the Dice coefficient, there is much to be improved. In this paper, we propose a modified u-net and curriculum learning strategy using a multi-task semi-supervised attention-based model, initially introduced by Chen et al., to segment ICH sub-groups from CT images. Using a modified inverse-sigmoid-based curriculum learning training strategy, we were able to stabilize Chen's algorithm experimentally. This semi-supervised model produced higher Dice coefficient values in comparison to a supervised counterpart, regardless of the amount of labeled data used to train the model. Specifically, when training with 80% of the ground truth data, our semi-supervised model produced a Dice coefficient of 0.67, which was higher than 0.61, obtained by a comparable supervised model. This result also surpassed by a greater margin the one obtained by using the out-of-the-box u-net by Hssayeni et al.

Keywords: segmentation; semi-supervised; ICH; hemorrhage; u-net; attention; curriculum

1. Introduction

The main risk factors of Intracranial Hemorrhage (ICH), which have extremely high rates of mortality, include hypertension and cerebral amyloid angiopathy. Other risk factors include alcohol intake, low levels of cholesterol, the $\epsilon 2$ and $\epsilon 4$ alleles of the Apolipoprotein E gene, anticoagulation treatment, and drug abuse. A depressing fact is that ICH occurs twice as frequently in low-/middle-income countries as in high-income countries [1]. Traditionally, ICH can be diagnosed from inspecting CT scans of the patient by a medical specialist. To save the time and effort of medical specialists, automated detection of ICH from CT scans becomes important. In addition, such a practice can improve the ICH diagnosis rate, especially in areas with limited access to medical professionals, potentially saving countless lives [2,3].

Deep learning has been applied to medical images for interpretation tasks with promising results [4]. For instance, Kervadec et al. [5] created a curriculum-style strategy for a semi-supervised CNN designed for segmentation tasks based on inequality constraints. This model was tested on left ventricle segmentation using MRI scans. However, it was unable to outperform a constrained CNN, which was pointed out in [6], although it was within a 1–3% margin when the number of labeled patients was increased to 40. Deep learning algorithms have further been used to find abnormalities in CT images of head [7] and chest [8], among others, allowing us to automate such tasks. Recently,

deep learning has found applications in hemorrhage detection. Many researchers have focused on solving this problem by either directly detecting ICH in general or a specific sub-group of ICH in a given image [9–11]. Most have used small datasets simply due to the limited availability of medical image data [12–14]. As regards the ICH sub-groups, they were classified as follows: Intraventricular (IVH), Intraparenchymal (IPH), Subarachnoid (SAH), Epidural (EDH), and Subdural (SDH). Others have worked on segmentation in order to highlight the specific regions where the ICH lies, assuming one was present [15,16]. Yuh et al. [17] used basic pattern recognition techniques in conjunction with a threshold-based algorithm in order to detect ICH, demonstrating 98% sensitivity with 59% specificity for ICH detection.

Shahangian et al. [18] proposed a hemorrhage detection algorithm using a variant of distance regularized level set evolution along with shape and texture features to detect and extract these regions. This method was deemed to work well with certain hemorrhage types, such as EDH, where obvious borders exist, in which case it was able to achieve a similarity rate above 75%. However, it preformed quite poorly with other types of hemorrhage, such as SDH, where it achieved below a 40% similarity rate. Kuo et al. [19] created a fully convolutional neural network based on the PatchFCN model [20] and trained on a dataset of approximately 4400 CT scans. Their model was then run on 200 test images, with the results being compared to those of four American Board of Radiology certified radiologists. It was observed that their model beat two of the said radiologists. These results were based on a binary decision as to whether or not an ICH was present in a given image with a segmentation task not in consideration. Another paper by Chilamkurthy et al. [4] introduced four algorithms for detecting sub-types of ICH trained on a large dataset containing nearly 300,000 CT scans. The average sensitivity was 92%, but the average specificity fell short at only 68% [17]. Much more recently, a paper by Hssayeni et al. detailed the use of a conventional u-net to detect ICH regions in CT scans. Their low Dice coefficient of 0.31 based on a five-fold cross-validation left room for improvement [2]. Cho et al. [21] introduced affinity graphs, an undirected weighted graph representing pixel connectivity, where classes with each affinity were defined with a segmentation mask and indicator function. This model was based on a traditional u-net architecture with an additional graph-based segmentation network following the output of the u-net. Their results produced a Dice score of 0.623, marginally beating a conventional u-net.

Previous work involving unsupervised image segmentation has proven to perform reasonably well compared to supervised alternatives, especially in the context of medical image segmentation [22]. A paper by Moriya et al. [23] proposed a deep representation learning approach of unsupervised segmentation clustering to combat the low amount of available labeled medical image data. This was done by learning deep feature representations of training patches from a given image with joint unsupervised learning. Other unsupervised techniques, including autoencoders [24], restricted Boltzmann machines [25], deep belief networks [26], deep Boltzmann machine [27], and generative adversarial networks [28], have been studied, but not many could reach the level achieved by traditional supervised learning techniques [29,30].

Semi-supervised learning has become a popular learning technique in recent years. Much analysis has been done on this technique and its usefulness [31]. With the scarcity of labeled medical data, the use of semi-supervised learning has become an attractive alternative. Researchers recently introduced a method called Dynamic Self-Training and Class-Balanced Curriculum (DST-CBC), specifically for semi-supervised semantic segmentation with the ability to exploit all unlabeled data by training with pseudo-labels. This approach was shown to beat narrowly other state-of-the-art models on the PASCAL VOC 2012 and Cityscapes datasets [32]. In some instances, semi-supervised learning techniques performed better than their supervised counterparts. For instance, Bortsova et al. [33] introduced a semi-supervised segmentation method that consistently learned under transformations, obtaining a higher segmentation accuracy than that of supervised learning.

In this paper, we introduce a modified u-net and curriculum learning strategy, using a semi-supervised model based on an earlier work by Chen et al. [34], to perform semantic segmentation

on a small dataset of ICH obtained from the Al Hilla Teaching Hospital in Iraq [2]. We also used the RSNA Intracranial Hemorrhage Detection dataset as a collection of unlabeled images to test our semi-supervised learning strategy. The newly-adopted curriculum learning technique augmented Chen's joint training strategy, which made the algorithm more stable. By employing a modified u-net in the encoder, we also reduced model overfitting due to over-parameterization.

The rest of this paper is organized as follows: Sections 2–4 introduce semi-supervised learning, the U-Netarchitecture, and attention, respectively. Section 5 details the experimental methods and procedures. Results are presented in Section 6, and concluding remarks are given in Section 7.

2. Semi-Supervised Learning

As has been mentioned before, in the field of automatic medical imaging, where it may involve the task of segmentation or classification, properly labeled data are difficult to obtain. For example, the RSNA Intracranial Hemorrhage Detection Dataset [35] required the collaboration of over four universities and more than 60 volunteers to label CT scans manually. Medical datasets set up for semantic segmentation training require even more resources, because professionally trained pathologists and radiologists need to draw the boundaries of hemorrhage regions manually for thousands of samples.

The difficulty in obtaining datasets large enough to take advantage of the innovations in the field of deep learning has inspired a new way of training, semi-supervised learning. Semi-supervised learning is able to use a large repertoire of unlabeled data along with a small number of labeled samples to create robust classifiers [36]. This is especially relevant to tasks involving semantic segmentation of medical images, where CT scan machines generate unlabeled data in the order of hundreds every hour. A properly trained classifier can label a sequence of CT images in a fraction of a second, while it may take a professionally trained radiologist a few minutes to label a single sample [37].

Formally, the goal of semi-supervised learning is to learn a function $f : X \rightarrow Y$ using a (potentially noisy) set of labeled data, $(x_1, y_1), ..., (x_n, y_n)$ and a set of unlabeled data $(x_{n+1}, ..., x_{n+k})$ subject to $n \ll k$ with the assumption that $\forall x_i \in X$. The principle advantage to semi-supervised learning boils down to improving generalization and reducing overfitting on the training set. By forcing the network to learn features corresponding to a larger dataset, we effectively regularized our network against non-generalizing local minima that existed within the training process. In this paper, we solved this problem by using an autoencoder with a modified u-net initially proposed by Chen et al. [34] on two separate, independent sources of data detailed in Section 5.1.

3. U-Net

The difficulty of procuring large, high-quality datasets revolving around medical imagery created the need for a new type of network, u-nets. These networks are well suited for segmentation tasks, which was proven by the network's victory while in its infancy in the EMsegmentation challenge at ISBI2012 [38]. Without any fully-connected layers, the resulting segmentation map simply consists of pixels for which a full context is revealed in a given image, allowing for the seamless segmentation of large images through an overlap-tile strategy. These techniques work for high resolution images that otherwise would be difficult to analyze due to current GPU memory limitations [39].

The architecture of u-nets is based on fully-convolutional networks, consisting of convolutional, deconvolutional, upsampling, and pooling layers. Each u-net is made up of a contracting encoder, which analyzes the entirety of the input image, and an expansive decoder, which produces a full-size segmentation [40,41]. More specifically, the encoder is a typical Convolutional Neural Network (CNN) with multiple convolutions followed by a Rectified Linear Unit (ReLU) and a 2×2 max pooling operation. It is followed by the decoder, which combines feature information through a 2×2 up-convolution [39]. Usually, these networks use 2D inputs and outputs, although other similar networks have been produced to operate on 3D data. One such network was described in a paper by Çiçek et al. [40].

The architecture of the u-net focuses on the following two objectives: (1) capturing and summarizing coarse-grained features and (2) using fine-grained information for inference. By adding a contracting and expanding autoencoder structure, the u-net can achieve the first objective despite its exponentially large receptive field (which grows by a factor of two per layer in the network). This allows the final layer of the network to have gradient access to a large window of the input and to compute a summary that is most informative for segmentation, which is the second objective. To take off the load of reconstruction, which many autoencoders treat as their objective function, and to leverage fine-scale information, residual connections of the u-net are then able to span across the encoder and decoder. This both allows the final layer to have access to the initial input and reduces the number of chain-rules required for gradient descent, which reduces the impact of the vanishing gradient problem.

4. Attention

An attention mechanism in a neural network allows the neural network to focus on specific features unique for a particular application. The attention mechanism creates a mask to weigh features extracted by the neural network, which increases the importance of certain features and reduces that of others [42]. A variety of attention mechanisms exist. Those suitable to our applications are soft and hard attention models and local attention models. Usually, both soft and hard attention models process images through a CNN and a Long Short-Term Memory (LSTM) network to extract features and produce descriptions. The main difference between these two types is that soft attention models use the entirety of the input image, while hard attention models only use a subsection of the input [43]. Local attention models essentially combine hard and soft attention models, first predicting the outcome with the entirety of the input image and then localizing it with only a portion of the input [44]. By purposefully reducing and boosting the importance of specific regions of the feature-space, attention mechanisms can be seen as a type of regulation that forces the network to learn informative "ways to look around, forward and backward," allowing for faster convergence and better results. Conventionally, this regulation is enforced using a softmax function, driving the network to pick at which specific regions to look. In this study, we employed multi-task attention by using a variant of soft attention to separate foreground and background elements during unsupervised training. The multi-task attention mechanism worked to inform a reconstruction task by generating weights from the segmentation network. This process is explained in greater detail in Section 5.2.

5. Methods

5.1. Dataset

In this study, our goal was to perform semantic segmentation of ICH using a small number of labeled elements and a large repertoire of unlabeled elements, as is the typical setup in real-life scenarios. Our labeled data points were CT scans obtained from 36 patients diagnosed with ICH and included the following types: IVH, IPH, SAH, EDH, and SDH. The data were obtained from the Al Hilla Teaching Hospital in Iraq and were collected between February and August 2018. Each CT scan for each patient included about 30 slices with a 5 mm slice thickness [2]. Out of the 36 diagnosed patients, there were on average about 9 slices for each CT scan that indicated hemorrhage among those patients, which had been annotated with ICH regions, totaling 318 256 $\times$ 256 ground truth images.

The unlabeled data used were from the Radiological Society of North America (RSNA) ICH dataset and contained around 25,000 CTs of patients diagnosed with ICH, totaling approximately 4,000,000 CT slices, of which approximately 250,000 were diagnosed positive for ICH. The RSNA dataset while diagnosed and labeled did not contain any segmentations. For our experiments, we randomly picked 100,000 samples or slices from 250,000 ICH positive samples, which were downsampled to 256 $\times$ 256. All images were provided in the DICOMformat with metadata containing multiple properties, allowing us to properly window the data to only look at the relevant features (brain window). The RSNA

recruited more than 60 volunteers to diagnose and classify more than 25,000 CT scans to assemble this dataset. The original de-identified CT studies were provided by Stanford University, Thomas Jefferson University, Unity Health Toronto, and Universidade Federal de São Paulo (UNIFESP) [35].

5.2. Model Architecture

In our study, we used a modified u-net architecture, as shown in Figure 1. In the early stage of our study, we found that using the full u-net on the limited data severely over-parameterized the network, leading to slow convergence and overfitting during training. To counter this, we reduced the number of layers for both the encoder and decoder. To compensate for the loss of encoder and decoder layers, we used transposed convolutions rather than upsampling blocks. Similar to the conventional u-net, each level of our network consisted of a convolution, a batch normalization, and then, a max pooling operation for the encoder and a transposed convolution and a batch normalization operation for the decoder. There were also residual concatenation connections that spanned across each level and the entire network, utilizing ReLUs for their activation function. Using this u-net, we ran three experiments: (1) pretraining the u-net's encoder on one of the larger classification datasets, then performing transfer learning on a smaller dataset for image segmentation; (2) performing supervised training using only labeled data; and (3) testing the proposed method of semi-supervised multi-task attention using both the larger unlabeled and smaller labeled datasets. For the first two tasks, we used the u-net shown in Figure 1. For the third task, the encoder and decoder structure was identical to Figure 1, except that the residual connections were removed for the second (unsupervised) decoder and that the output included two feature maps rather than one [34].

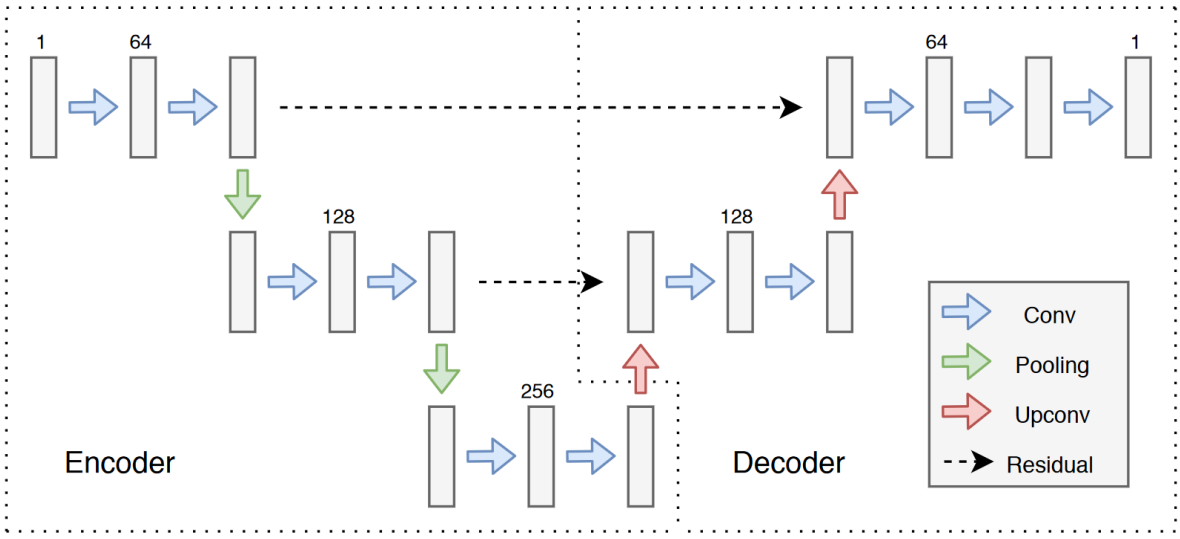

Figure 1. Architecture of our modified u-net used in our models. Each convolution (blue) has a kernel size of 3 × 3 and a padding of 1 to retain the image size. The pooling operation (green) has a kernel size of 2 × 2 and a stride of 2, and similarly, the upconv (red) (or transposed convolution operator) also has a kernel size of 2 × 2 with a stride of 2. The number of feature maps for each CNN block is noted above the block. This network is used as both the baseline pretrained model and the supervised model. The semi-supervised model has a similar network architecture except that the second autoencoder decoder does not have residual connections to the encoder, and the final output consists of two feature maps, rather than one.

For the semi-supervised learning problem, previous solutions have involved pretraining, proxy labeling, or proxy learning [45]. In this study, rather than directly training on the predicted segmentation labels generated by our model on unlabeled data, we instead used a multi-task attention mechanism to separate the foreground and background of the input image, and then tasked the

unsupervised autoencoder to reconstruct both the foreground and the background. The supervised and unsupervised portions of the network were trained in a schedule alteration, which is described in more detail in Section 5.3. The model architecture is shown in Figure 2. Imposing multi-task reconstruction allowed the encoder to learn a wider range of features, which may not be present in a very limited training set. This would in turn improve the accuracy performance of the model even though it was trained with a limited number of labeled samples.

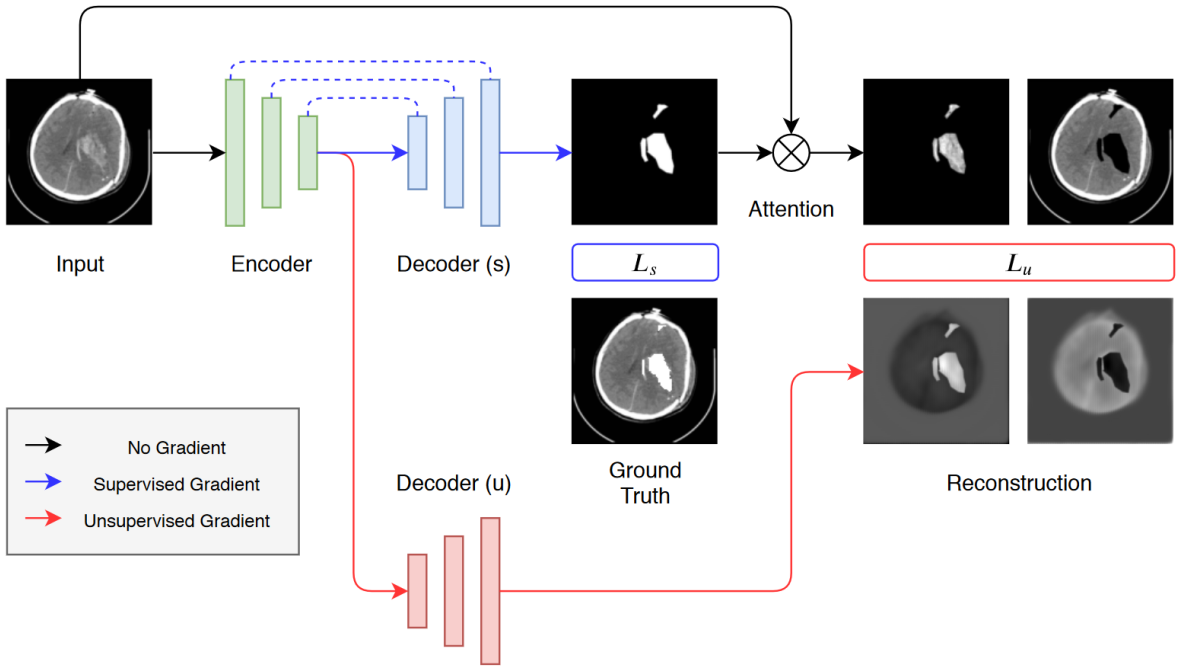

Figure 2. Semi-supervised attention network with one shared encoder (green) and two decoders. The first one (blue) is the same decoder shown in Figure 1, and the second one (red) is the same decoder except without residual connections (effectively making it an autoencoder when paired with the encoder). The color of an arrow signifies the gradient flow between the supervised and unsupervised portions of the model. The loss functions L_s and L_u correspond to the supervised and unsupervised loss, respectively. The training procedure, gradient flow, and loss functions are detailed in Section 5.3.

5.3. Training Procedure and Loss Functions

While Chen et al. [34] used both joint and alternating training strategies, after experimentation, we found that both strategies (especially when dealing with a very limited and volatile dataset) led to unstable behavior, causing the loss to explode early on in training. We instead proposed a curriculum learning strategy, which decayed with probability as training progressed. This mechanism regulated alternating learning, which is explained below.

Originally, Chen's alternating learning strategy optimized alternately the supervised portion of the network and the unsupervised portion of the network. In each run, the unsupervised training randomly picked data points from the larger dataset. For example, training the network with a labeled dataset of 100 points and an unlabeled dataset of 1000 points would proceed with first training the model's supervised portion with the 100 labeled points, then training the unsupervised portion with 100 randomly selected samples from the 1000 unlabeled data points. Through the experimental study, we found this method could lead to gradient explosions. We hypothesized that if the supervised portion of the network were inaccurate, the attention mechanism would identify the wrong portions of the image for the autoencoder to reconstruct. This pushed the weights of the autoencoder to change in the wrong direction, resulting in divergence.

To solve the encountered problem, we proposed to use an inverse sigmoid curriculum learning strategy, which is outlined next. Rather than alternating equally between training the supervised and unsupervised portion of the network, we decided which portion to train based on a Bernoulli random variable with parameter p, where $p = f(x)$ and x is the current epoch. Let $f(x) = \max\left(\frac{k}{k+\exp(\frac{x}{k})}, 0.5\right)$, where k is a preset hyperparameter. This allowed the algorithm at the start of the training to optimize mainly the supervised portion of the model, then eventually converge to Chen's alternating training algorithm. This version of the inverse-sigmoid learning curriculum was first introduced by Bengio et al. [46] when training forecasting systems. We found in our experimental study that this curriculum learning strategy produced better results, which will be shown later, in comparison to other training methods.

The loss functions L_s and L_u, shown in Figure 2, are the Jaccard loss and Weighted Mean Squared Error (WMSE), respectively. Formally, the L_u loss is defined in Equation (1), which weighs the MSE of the foreground and background reconstruction by the size of the segmentation as follows:

$$L_u = \frac{\sum \tilde{y}_b^{(i)}}{N} \text{MSE}[x \odot \tilde{y}_b, \hat{y}_b] + \frac{\sum \tilde{y}_f^{(i)}}{N} \text{MSE}[x \odot \tilde{y}_f, \hat{y}_f] \tag{1}$$

where $\tilde{y}$ and $\hat{y}$ are the predictions of the reconstruction and segmentation paths, respectively, for the background (b) and foreground (f); N is the number of voxels in an input image x; and $\odot$ is the Hadamard product [34].

To show the effectiveness of our semi-supervised model, we trained and tested our model with N-fold cross-validation, where the training set contained 20%, 50%, and 80% of the labeled data, corresponding to the 5-fold, 2-fold, and 5-fold cross-validation schemes, respectively. In an N-fold cross-validation procedure, also called the Leave-One-Person-Out (LOPO) procedure, the dataset is divided into N folds, in which N-1folds are used for training and one for validation. This procedure is repeated N times until all of the folds are used once for validation. Because our dataset was relatively small, our model was trained with the Limited Broyden–Fletcher–Goldfarb–Shannon (L-BFGS) algorithm, a quasi-Newtonian optimizer, for the supervised portion and the Adam optimizer for the unsupervised version. We used a learning rate of 10^{-4}, trained for 4000 epochs, and set k in our decay function to be 40.

5.4. Performance Metrics

To measure the performance of our model, we used both the Sørensen–Dice coefficient (or simply Dice coefficient) and Jaccard coefficient, shown in Equations (2) and (3), respectively.

$$Dice = 2 * \frac{|\hat{y} \cap y|}{|\hat{y}| + |y|} \tag{2}$$

$$Jaccard = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|} \tag{3}$$

where $\hat{y}$ and y represent the predicted and ground truth, respectively. In our case, y and $\hat{y}$ are the observed and predicted segmentation regions of ICH, respectively.

Both metrics, which range from 0 to 1, gauge the similarity, or overlap, of two sets; however, unlike the Dice coefficient, the Jaccard coefficient satisfies the triangular inequality and is, therefore, a proper distance measure. This is why it is preferred to optimize the Jaccard coefficient rather than the Dice coefficient, though they often produce similar results.

6. Results

In the experimental study, we evaluated the performance of the trained models in terms of both Dice and Jaccard coefficients and averaged the results for each individual run during cross-validation. The results, which are given in Table 1, show that our semi-supervised model beat all other models

tested, independently of the amount of data used to train them. However, it was interesting to see that the margin of performance gain decreased between the supervised and semi-supervised models as the amount of data used to train increased. This was because the semi-supervised methodology promoted feature learning through the reconstruction of the background and foreground after attention was introduced during a segmentation task.

While Hssayeni et al. was the only other research team to train and test on the same dataset as our team; other teams have developed ICH segmentation models that operated on other datasets. In Table 2, we show the Dice (Jaccard) coefficients of our best performing model compared with some other models and the amount of training data used. While Chang et al. [9] were able to get higher Dice and Jaccard coefficients than us, they did so with 160 times the training samples, which was an expected result due to the universal approximation theorem.

Table 1. Dice (Jaccard) coefficients obtained by various methods trained with 20%, 50%, and 80% of ground truth data. The first row corresponds to the u-net employed in Hssayeni et al.; the second row corresponds to Chen et al.'s algorithm applied directly on our two datasets; the third row corresponds to our modified u-net pretrained on the classification task from the RSNA dataset; the fourth row corresponds to our modified u-net trained only in a supervised fashion; and the fifth row corresponds to our model combining the modified u-net and unsupervised attention autoencoder with curriculum learning.

Data Used	20%	50%	80%
Hssayeni et al. [2]	-	-	0.31 (0.18)
Chen et al. [34]	Diverged	Diverged	Diverged
Pretrained	0.33 (0.20)	0.44 (0.28)	0.61 (0.44)
Supervised	0.31 (0.18)	0.42 (0.27)	0.61 (0.44)
Semi-Supervised	**0.44 (0.28)**	**0.51 (0.34)**	**0.67 (0.50)**

Table 2. Dice (Jaccard) coefficients obtained by various methods for semantic segmentation of ICHs along with the number of training samples used. Chang et al. and Kuang et al. used different datasets from the one used by Hssayeni et al. and us. Most notably, our algorithm is shown to beat Kuang et al.'s algorithm despite only having access to a third of the labeled data. † Chang's dataset composed of approximately 901 patients totaling approximately 40,000 images containing ICH. ‡ Kuang's dataset comprised of approximately 90 patients. With approximately 80 depth slices per 3D CT sample, we estimate the amount of information contained within his dataset to be equivalent to approximately 720 data points.

Method	Training Samples	Dice (Jaccard)
Hssayeni et al. [2]	254	0.31 (0.18)
Chang et al. [9]	40,000 †	0.85 (0.74)
Kuang et al. [14]	720 ‡	0.65 (0.48)
Cho et al. [21]	6000	0.62 (0.45)
Proposed	254	0.67 (0.50)

To help us to appreciate the effectiveness of the proposed model, Figure 3 shows some examples of the ICH segmentation across the performance spectrum. The results were obtained by our semi-supervised model with 80% of data for training.

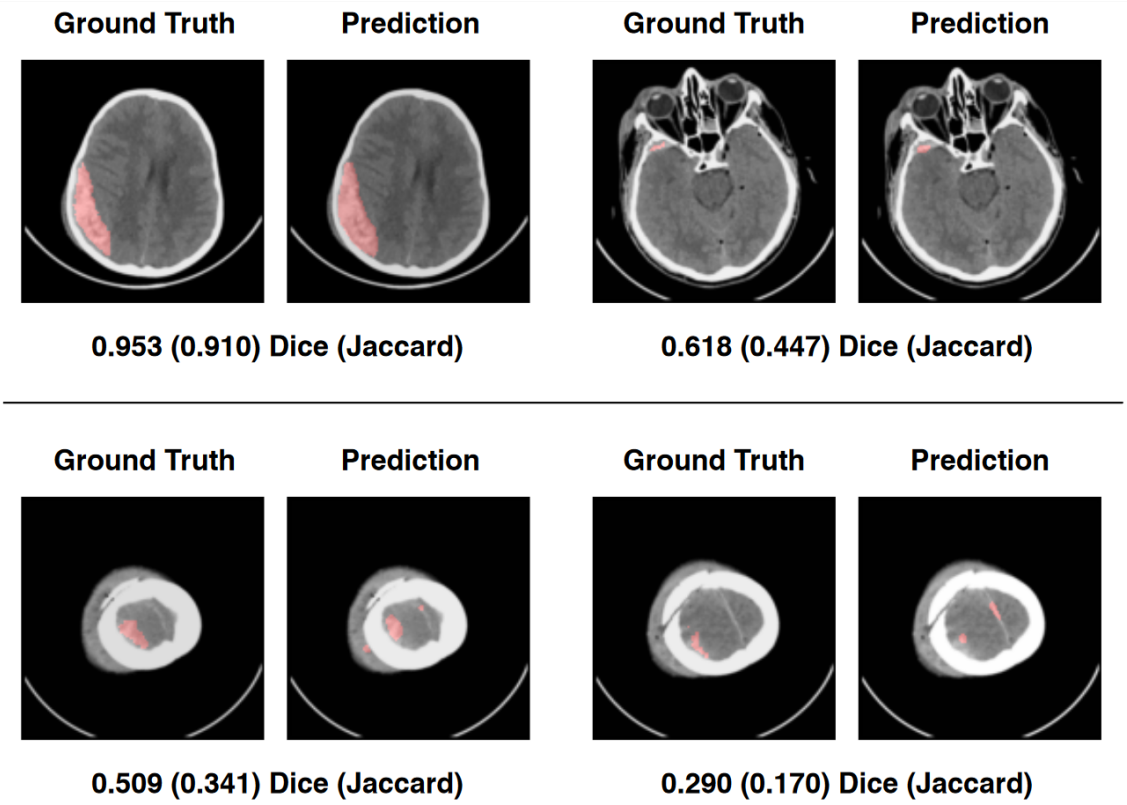

Figure 3. Ground truth and prediction pulled from our segmentation model (semi-supervised trained on 80% of our data). The corresponding Dice (Jaccard) coefficient values are also given. Purple regions indicate intracranial hemorrhages. Please be advised that optimizing either Dice or Jaccard coefficients produced almost identical segmentation results visually; therefore, only one predicted image for each example is shown.

7. Conclusions

Intracranial Hemorrhage (ICH) is a severe condition with extremely high rates of mortality. Identification and segmentation of CT scans of patients suspected of suffering ICH are vital to formulating treatment and surgery plans. Despite the importance of this issue, there are very few reliable solutions to ICH segmentation without a medical professional. Therefore, it is imperative to design accurate and robust methods to segment ICH areas from CT scans. With high enough accuracy rates, these models could potentially outperform trained professionals, leading to fewer false-negative ICH detections. Unfortunately, due to the cost and nature of acquiring expert-labeled CT scans of ICH patients, the repository for training data for newer deep learning algorithms is often not enough to produce robust segmentation models.

In this paper, we proposed a modified u-net and curriculum learning strategy for the semi-supervised model initially introduced by Chen et al. [34] to segment ICH regions from patient CT scans automatically. The adopted curriculum learning strategy solved the gradient explosion problem that was encountered during our experiments of Chen's alternate learning method. The central idea of this training procedure was to optimize mainly the supervised portion of the model at the beginning, then eventually converge to Chen's alternating learning algorithm. This new model worked with a small labeled dataset and a large unlabeled dataset. With our segmentation model, we trained and tested a purely supervised version and pretrained modified u-net and showed that our model surpassed both regardless of the amount of data used to train it.

This being said, the Dice and Jaccard coefficients of our final solutions were still far from perfect. Due to the volatile nature of performing machine learning on small datasets, we could not guarantee the integrity of this algorithm when extended to different domains and datasets. While we could

empirically show that our algorithm had an improved stability compared to the original method proposed by Chen et al. [34], more work and experimentation need to be done with a variety of tasks and datasets to confirm the reliability of its results.

In the future research, we plan to study ways to improve the accuracy and robustness of the segmentation model. As seen in Figure 3, one issue that our model encountered was the inability to segment smaller regions while optimizing larger ones. This is a common problem in semantic segmentation and is mainly due to class imbalance. A future direction is to integrate the semi-supervised model with other loss functions, say Tversky loss [47], to combat class imbalance.

Author Contributions: Conceptualization, J.L.W.; methodology, J.L.W.; software, J.L.W.; validation, J.L.W.; formal analysis, J.L.W.; investigation, J.L.W.; resources, H.Z.; data curation, A.K.I.; writing, original draft preparation, J.L.W. and H.F.; writing, review and editing, J.L.W., H.F., and H.Z.; visualization, J.L.W.; supervision, H.Z.; project administration, H.Z. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ICH	Intracranial Hemorrhage
IVH	Intraventricular Hemorrhage
IPH	Intraparenchymal Hemorrhage
SAH	Subarachnoid Hemorrhage
EDH	Epidural Hemorrhage
SDH	Subdural Hemorrhage
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
LSTM	Long Short-Term Memory network
RSNA	Radiological Society of North America
WMSE	Weighted Mean Squared Error

References

1. Caceres, A.J.; Goldstein, J.N. Intracranial Hemorrhage. *Emerg. Med. Clin. N. Am.* **2012**. [CrossRef]
2. Hssayeni, M.D.; Croock, M.S.; Al-Ani, A.; Al-khafaji, H.F.; Yahya, Z.A.; Ghoraani, B. Intracranial Hemorrhage Segmentation Using Deep Convolutional Model. *arXiv* **2019**, arXiv:1910.08643.
3. Litjens, G.J.S.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
4. Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N.G.; Venugopal, V.K.; Mahajan, V.; Rao, P.; Warier, P. Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head CT Scans. *arXiv* **2018**, arXiv:1803.05854.
5. Kervadec, H.; Dolz, J.; Granger, E.; Ayed, I.B. Curriculum semi-supervised segmentation. *arXiv* **2019**, arXiv:1904.05236.
6. Kervadec, H.; Dolz, J.; Tang, M.; Granger, E.; Boykov, Y.; Ayed, I.B. Constrained-CNN losses forweakly supervised segmentation. *Med. Image Anal.* **2018**, *54*, 88–99. [CrossRef]
7. Gao, X.; Hui, R.; Tian, Z. Classification of CT brain images based on deep learning networks. *Comput. Methods Programs Biomed.* **2016**, *138*. [CrossRef]
8. Anthimopoulos, M.; Christodoulidis, S.; Ebner, L.; Christe, A.; Mougiakakou, S. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Trans. Med. Imaging* **2016**, *35*, 1207–1216. [CrossRef]
9. 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. *Am. J. Neuroradiol.* **2018**, *39*, 1609–1616. [CrossRef]

10. Ye, H.; Gao, F.; Yin, Y.; Guo, D.; Zhao, P.; Lu, Y.; Wang, X.; Bai, J.; Cao, K.; Song, Q.; et al. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *Eur. Radiol.* **2019**, *29*, 6191–6201. [CrossRef]
11. Bar, A.; Mauda, M.; Turner, Y.; Safadi, M.; Elnekave, E. Improved ICH classification using task-dependent learning. *arXiv* **2019**, arXiv:1907.00148.
12. Li, Y.; Wu, J.; Li, H.; Li, D.; Du, X.; Chen, Z.; Jia, F.; Hu, Q. Automatic Detection of the Existence of Subarachnoid Hemorrhage from Clinical CT Images. *J. Med. Syst.* **2012**, *36*, 1259–1270. [CrossRef] [PubMed]
13. Lee, H.; Yune, S.; Mansouri, M.; Kim, M.; Tajmir, S.H.; Guerrier, C.E.; Ebert, S.A.; Pomerantz, S.R.; Romero, J.M.; Kamalian, S.; et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat. Biomed. Eng.* **2018**, *3*, 173–182. [CrossRef] [PubMed]
14. Kuang, H.; Menon, B.K.; Qiu, W. Segmenting Hemorrhagic and Ischemic Infarct Simultaneously From Follow-Up Non-Contrast CT Images in Patients with Acute Ischemic Stroke. *IEEE Access* **2019**, *7*, 39842–39851. [CrossRef]
15. Prakash, K.N.B.; Zhou, S.; Morgan, T.C.; Hanley, D.F.; Nowinski, W.L. Segmentation and quantification of intra-ventricular/cerebral hemorrhage in CT scans by modified distance regularized level set evolution technique. *Int. J. Comput. Assist. Radiol. Surg.* **2012**, *7*, 785–798. [CrossRef] [PubMed]
16. Kuo, W.; Häne, C.; Yuh, E.L.; Mukherjee, P.; Malik, J. Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection. *arXiv* **2018**, arXiv:1809.02882.
17. Yuh, E.L.; Gean, A.D.; Manley, G.T.; Callen, A.L.; Wintermark, M. Computer-Aided Assessment of Head Computed Tomography (CT) Studies in Patients with Suspected Traumatic Brain Injury. *J. Neurotrauma* **2008**, *25*, 1163–1172. [CrossRef]
18. Shahangian, B.; Pourghassem, H. Automatic brain hemorrhage segmentation and classification algorithm based on weighted grayscale histogram feature in a hierarchical classification structure. *Biocybern. Biomed. Eng.* **2015**, *36*. [CrossRef]
19. Kuo, W.; Häne, C.; Mukherjee, P.; Malik, J.; Yuh, E. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 201908021. [CrossRef]
20. Kuo, W.; Häne, C.; Yuh, E.L.; Mukherjee, P.; Malik, J. PatchFCN for Intracranial Hemorrhage Detection. *arXiv* **2018**, arXiv:1806.03265.
21. Cho, J.; Choi, I.; Kim, J.; Jeong, S.; Lee, Y.S.; Park, J.; Kim, J.; Lee, M. Affinity Graph Based End-to-End Deep Convolutional Networks for CT Hemorrhage Segmentation. In Proceedings of the 2019 International Conference on Neural Information Processing, Sydney, Australia, 12–15 December 2019; pp. 546–555. [CrossRef]
22. Sivanesan, U.; Braga, L.H.; Sonnadara, R.R.; Dhindsa, K. Unsupervised Medical Image Segmentation with Adversarial Networks: From Edge Diagrams to Segmentation Maps. *arXiv* **2019**, arXiv:1911.05140.
23. Moriya, T.; Roth, H.R.; Nakamura, S.; Oda, H.; Nagara, K.; Oda, M.; Mori, K. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. *arXiv* **2018**, arXiv:1804.03830. doi:10.1117/12.2293414.
24. Bourlard, H.; Kamp, Y. Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biol. Cybern.* **1988**, *59*, 291–294. [CrossRef] [PubMed]
25. Montúfar, G. Restricted Boltzmann Machines: Introduction and Review. *arXiv* **2018**, arXiv:1806.07066.
26. Koo, J.; Klabjan, D. Improved Classification Based on Deep Belief Networks. *arXiv* **2018**, arXiv:1804.09812.
27. Salakhutdinov, R.; Hinton, G. Deep Boltzmann Machines. In Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Clearwater, FL, USA, 16–19 April 2009; van Dyk, D., Welling, M., Eds.; PMLR: Clearwater Beach, FL, USA, 2009; Volume 5, pp. 448–455.
28. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Nice, France, 2014; pp. 2672–2680.
29. Raza, K.; Singh, N.K. A Tour of Unsupervised Deep Learning for Medical Image Analysis. *arXiv* **2018**, arXiv:1812.07715.
30. Luc, P.; Couprie, C.; Chintala, S.; Verbeek, J. Semantic Segmentation using Adversarial Networks. *arXiv* **2016**, arXiv:1611.08408.

31. Cheplygina, V.; de Bruijne, M.; Pluim, J.P.W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *arXiv* **2018**, arXiv:1804.06353.
32. Feng, Z.; Zhou, Q.; Cheng, G.; Tan, X.; Shi, J.; Ma, L. Semi-Supervised Semantic Segmentation via Dynamic Self-Training and Class-Balanced Curriculum. *arXiv* **2020**, arXiv:2004.08514.
33. Bortsova, G.; Dubost, F.; Hogeweg, L.; Katramados, I.; de Bruijne, M. Semi-Supervised Medical Image Segmentation via Learning Consistency under Transformations. *arXiv* **2019**, arXiv:1911.01218.
34. Chen, S.; Bortsova, G.; Juarez, A.G.; van Tulder, G.; de Bruijne, M. Multi-Task Attention-Based Semi-Supervised Learning for Medical Image Segmentation. *arXiv* **2019**, arXiv:1907.12303.
35. RSNA Intracranial Hemorrhage Detection. Available online: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection (accessed on 6 December 2019).
36. Liu, X.; Zachariah, D.; Wågberg, J.; Schön, T.B. Reliable Semi-Supervised Learning when Labels are Missing at Random. *arXiv* **2018**, arXiv:1811.10947.
37. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.; Wu, Z.; Ding, X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *arXiv* **2019**, arXiv:1908.10454.
38. Ciresan, D.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Nice, France, 2012; pp. 2843–2851.
39. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
40. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv* **2016**, arXiv:1606.06650.
41. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038.
42. Park, D.H.; Hendricks, L.A.; Akata, Z.; Schiele, B.; Darrell, T.; Rohrbach, M. Attentive Explanations: Justifying Decisions and Pointing to the Evidence. *arXiv* **2016**, arXiv:1612.04757.
43. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
44. Luong, M.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
45. Tran, P.V. Exploring Self-Supervised Regularization for Supervised and Semi-Supervised Learning. *arXiv* **2019**, arXiv:1906.10343.
46. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. *arXiv* **2015**, arXiv:1506.03099.
47. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *arXiv* **2017**, arXiv:1706.05721.