Classification of Neonatal Brain Ultrasound Scans Using Deep Convolutional Neural Networks

Dongwoon Hyun[1], Leandra Brickson[2]
Departments of [1]Bioengineering and [2]Electrical Engineering,
Stanford University, Stanford, CA 94305
{dhyun,llbricks}@stanford.edu

I. Introduction

Neonatal neurosonography (NS) is a non-invasive medical imaging technique used to evaluate the neonatal or infant brain for abnormalities (e.g., hydrocephalus, hemorrhages, infections) using ultrasonography[1]. While ultrasound imaging of the brain is difficult in adults because of the lack of an acoustic window, neonates possess an opening at the top of the skull, specifically the *anterior fontanelle*, through which NS may be performed. NS provides physicians with real-time diagnostic information that can be used to guide treatment. The automatic annotation and captioning of radiological images has the potential to further aid physicians in making fast and accurate diagnoses. Recent work has shown that radiological images can be annotated by using a combination of natural language processing (NLP) analysis of the textual radiological reports and deep convolutional neural network (CNN) analysis of the images [2]. In this work, we aim to achieve similar success in classifying and/or annotating neonatal NS images. This project is composed of two parts: image classification using features from textual reports on a set of 2D NS images with supervised learning, and the automatic labeling of text reports using unsupervised learning NLP techniques. **The input to the image classification task is a NS image, and the output is a binary classification of either "normal" and "not normal". The input to the automated labeling task is a new physician's report, and the output is a binary classification of 0 or 1 for each of 21 different keywords.**

II. Related Work

Several approaches have been proposed to combine NLP with CNNs for the automatic generation of image captions and descriptions of everyday objects. A recent model by Vinyals et al. [3] utilizes CNNs to embed an input image in a vector space, followed by a recurrent neural network (RNN) to decode the fixed-length vector into a complete sentence. Other similar approaches involve a multimodal model utilizing a combined embedding of images and text [4], [5]. Each of these works focus on annotating everyday images that are readily available and can be labeled with little to no expertise, permitting a rich embedding of words and images.

By contrast, radiological images must be annotated by physicians who receive years of highly specialized training, making labeled data sets difficult to obtain. In a recent study, Shin et al. [2] successfully demonstrated a more limited form of annotation (keywords as opposed to complete sentences) of a large-scale ($\sim$200,000 image) radiological database using an approach similar to that of [3]. Unsupervised NLP techniques were used to obtain semantic labels, to model documents as a combination of latent topics, to learn a word-to-vector embedding, and to classify images. CNNs were then trained via regression to learn an image-to-vector mapping. The images were then annotated using words associated with the topics the nearest neighbors in the vector space.

Our work began with an approach similar to that of [2], in which we planned to annotate NS images by embedding the images and their associated text reports. However, due to time constraints, this work instead investigates the direct classification of NS images with various CNN architectures, as well as the reliability of automatically generated labels from text reports, which can later be used to create larger labeled training sets when additional images are added to the data set.

III. Dataset and Features

A. Neurosonography Report Data

With the assistance of pediatric radiologist Dr. Safwan Halabi, the de-identified medical reports of $N = 2372$ NS patients were obtained from the Lucille Packard Children's Hospital (LPCH) picture archiving and communication system (PACS). To expand the vocabulary and obtain a richer embedding of medical terminology, an additional 996 textual reports of head ultrasounds were obtained from the publicly available Open-i collection of biomedical images [6]. The entire corpus contained a combined total of 242397 words (after pre-processing), with a vocabulary of $V = 1543$ unique words that occurred 10 or more times.

B. Neurosonography Images

For 333 of the 2372 medical reports, a total of 11,205 NS head images and 3,232 videos were obtained from the LPCH PACS in DICOM format. Each of these files were de-identified according to HIPAA standards on a protected network by removing the top of the image and removing protected health information from the

Fig. 1: Example of a de-identified saggital view NS image.

DICOM header. An example de-identified NS image is shown in Fig. 1.

IV. Methods

A. Text Preprocessing

Several steps of preprocessing were applied on the raw text. First, the words were stemmed using a standard stemming algorithm called Porter2, as implemented in the `stemming` Python module. Next, a combination of the Sampling and Apriori algorithms [7] was used to find common recurring words and phrases of arbitrary length in the report text. This method was used to recursively find words and multi-word phrases with occurrences above a minimum threshold. Some of the resulting words and phrases are presented in Table I, along with their number of occurrences. Common multi-word phrases were converted into single-word representations using underscores (e.g., "INTRACRANIAL_HEMORRHAGE") in order to retain the meaning of a phrase, which may be lost when considering the words individually (e.g., "INTRACRANIAL HEMORRHAGE" versus "INTRACRANIAL" and "HEMORRHAGE" separately).

B. Image Preprocessing

When obtained from the PACS, many of the images were labeled with the anatomical orientation burned into the images themselves (as seen in Fig. 1, where 'COR' refers to the coronal anatomical plane), but not in the header data itself. Because the images are 2D cross-sections (as opposed to 3D volumes), this information is vital for determining the orientation of the brain relative to the transducer, and therefore an important feature for image segmentation. This text was extracted from the images by thresholding at known text brightness values and then applying text recognition via optical character recognition (OCR) to recognize common measurement orientations. Each image was also of variable size; to create a constant input image size, the smaller images were zero padded or cropped to be the same size. Finally, all images were normalized to have a max value of 1.

Phrase	Occurrences
ABOVE	926
HEMORRHAGE	568
ULTRASOUND	552
HEAD ULTRASOUND	475
NORMAL HEAD ULTRASOUND	335
LATERAL	320
LEFT	242
EVIDENCE OF	229
INTRACRANIAL HEMORRHAGE	211
VENTRICULOMEGALY	203

TABLE I: Top 10 medically relevant phrases in reports

C. Word-to-vector embedding

Words and phrases with similar meaning and/or appearing in similar contexts can be determined by using a word-to-vector embedding such that the vector representations of similar words are closely spaced. The preprocessed corpus was mapped into a $\mathbf{R}^{200}$ vector space using a word-to-vector embedding algorithm that is publicly available from TensorFlow [8]. The default hyperparameter settings provided by TensorFlow (initial learning rate of 0.025, window size of 10 words, 15 training epochs, etc.) were used. This work was not directly used in our eventual architecture, but was useful for generating a vocabulary and obtaining a set of keywords to guide manual classification of the images.

D. Manual Image Classification

The 333 text reports associated with the obtained images were manually labeled using a total of $K = 21$ keywords, some of which were learned via the word-to-vector embedding described above. Labeling was performed by 3 non-experts. Examples of keywords included "normal", "diffuse hypoxic ischemic encephalopathy", "extra-axial fluid", "germinal matrix hemorrhage", and "ventriculomegaly". For each keyword, every text report was classified based on whether it pertained to the keyword, yielding a classification matrix of size $y_{\text{all}} \in \{0,1\}^{333\times K}$. For the binary classification task of determining "normal" vs. "abnormal", only the "normal" keyword column of the matrix was used for a ground truth vector of $y_{\text{normal}} \in \{0,1\}^{333}$.

E. Deep Convolutional Neural Networks Architecture

The CNNs designed in this project were coded in Python using Keras with a Theano backend. Training was done on an NVIDIA TITAN X GPU on Dr. Sebastian Thrun's cluster using a 30,000 image training set and a 5,000 image test set, separated by patient

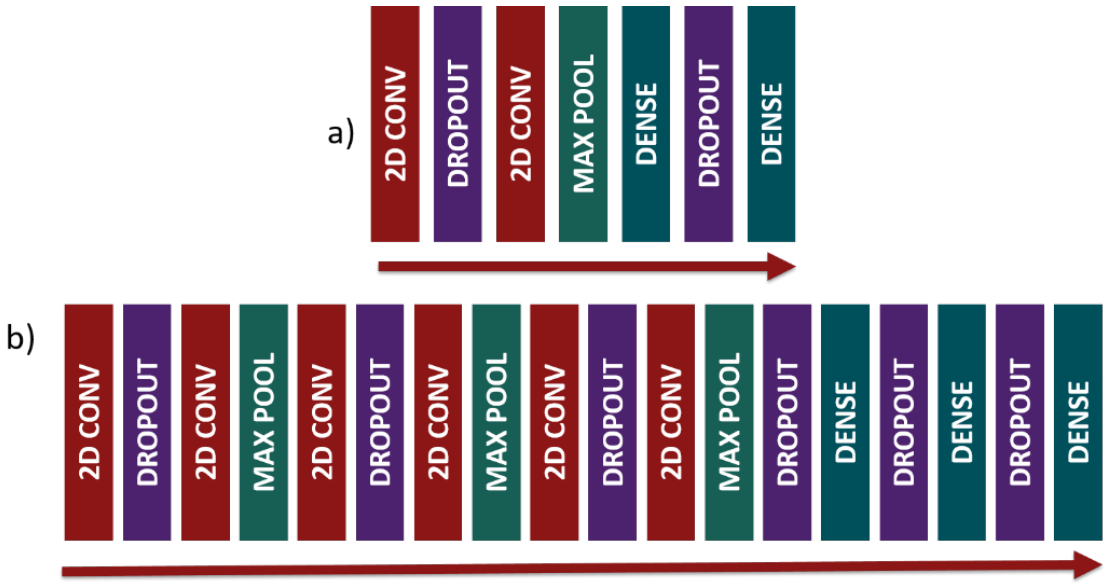

Fig. 2: a) Basic CNN and b) Deep CNN design architecture used for initial training.

report number. All images used were of the coronal orientation. For the architectures used, the large image size (700x700 pixels) created very large weight matrices, leading to memory allocation issues. To decrease the number of parameters, most networks had to be carefully constructed to downsample quickly via large maxpool layers or before network training.

Four different CNN designs were trained and tested. First, a shallow network shown in Fig. 2a was made to test basic classification. A deeper network, shown in Fig. 2b was then made to improve accuracy. Throughout the design of this network, many adjustments were tested, including changing the stride of the CONV layers, number of depth layers and pooling size to test what helped with classification. In addition to this, the conventional networks of AlexNet[9] and VGG16[10] were also implemented. For these final two designs, inputs were downsampled to 250x250 pixels before training.

Due to timing constraints on the project, each network was trained using only 5 epochs (With the exception of AlexNet, which was run for 20 epochs due to exceptional training speed). Ideally, each network would be trained with 20 or more epochs, but these trainings were enough to make an initial performance comparison.

F. Automatic Labeling of Text Reports

The manual labeling of text reports is non-trivial and time consuming, and will quickly grow infeasible moving forward as additional NS images and reports are added to our data set. To alleviate this burden, we investigated an NLP approach to automatically label new text reports. We used latent Dirichlet allocation[11] (LDA), an unsupervised learning algorithm, to discover an underlying set of topics from the collection of reports. LDA is a three-layer hierarchical Bayesian model that models each document as a mixture of T underlying topics. LDA is an appropriate model for physicians' reports because each report may be associated with one or more underlying conditions, diagnoses or other latent variable.

A trained LDA model takes a document-term vector $x \in \{0,1\}^V$ as input and returns $p \in [0,1]^T$, a probability distribution over the learned topics. We represent this as $f : x \mapsto p$. LDA is a mixture model, meaning that it is possible for a document to pertain to multiple topics. To associate the document with only the most relevant topics, we also define a thresholding function $g_\phi(p)$, such that

$$(g_\phi(p))_i = \begin{cases} 1 & p_i \geq \phi \\ 0 & p_i < \phi \end{cases}, \quad \text{for } i \in \{1, \ldots, T\}. \tag{1}$$

We then define a mapping from the thresholded topics to the labels, $h : g_\phi \mapsto y$ in order to generate labels from the topics. This allows us to input a document x_i and output a set of labels $\hat{y}_i \in \{0,1\}^{1\times K}$ as

$$\hat{y}_i = h(g_\phi(f(x_i))). \tag{2}$$

The training set $X_{\text{train}} \in \{0,1\}^{2039\times V}$ was selected as the document-term matrix of the $m_{\text{train}} = 2039$ textual reports for which we had no images. The test set $X_{\text{test}} \in \{0,1\}^{333\times V}$ was selected as the document-term matrix of the $m_{\text{test}} = 333$ reports which we had previously labeled in Section IV-D. The f mapping was first trained with an LDA model using the `lda` implementation available in the scikit-learn Python toolbox. (A total of $T = 10$ topics was found to give a reasonable balance between flexibility and minimizing repeated topics.) The resulting topics were then manually labeled in the same manner as the textual reports based on the top 15 words per topic to generate the h mapping. The threshold value was modulated from $\phi = 0$ to 1 in small increments, and the false-positive ($y = 0, \hat{y} = 1$) and true-positive ($y = 1, \hat{y} = 1$, also called precision) rates were measured to generate receiver operating characteristics (ROC) curves. The

area under the ROC curve (AUROC) was used as a quantitative metric. Additionally, accuracy was measured as

$$\text{Accuracy} = \frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total observations}}. \quad (3)$$

Quantitative assessment was performed both on the single-label classification task (normal vs. abnormal), as well as on a multi-label classification task in which the overall classification success rate is reported.

V. Results and Discussion

A. Image Classification

CNN Architecture	Accuracy	# Epochs
Basic CNN	59.0%	5
Deep CNN	65.6%	5
AlexNet	76.6%	20
VGG16	76.4 %	5

TABLE II: Summary of CNN Performance

The coronal image test set was manually classified as 65% normal and 35% abnormal. If we set a baseline to always classify the input as normal, then the baseline will be 65% accurate. Our goal with the neural network is then to classify with accuracy higher than this. Table II summarizes the accuracy values of each CNN architecture, which is determined by the percent of test samples misclassified on the network. The Basic and Deep CNN designs had accuracy at or lower than baseline, making them unsuitable for classification. However the AlexNet and VGG16 networks performed very well, at 76.6% and 76.4% accuracy, respectively.

B. Automatic Labeling of Text Reports

Topic #	Top Keywords
0	bilater small hemorrhag periventricular
5	region note lesion x mass
7	resist increas indic cerebral_arteri intracrani
8	normal head_ultrasound unremark
9	echogen find increas white_matt ischem

TABLE III: Top keywords for selected topics

Qualitatively, the LDA model did remarkably well at identifying different topics. A list of the top keywords for several topics generated from the training set are provided in Table III. Topic 0 was found to be relevant to intracranial hemorrhaging, topic 5 to lesions and masses, topic 7 to increased intracranial pressure, which increases the resistive index, topic 8 to normal patients, and topic 9 to diffuse ischemic hypoxia, which presents with an increased echogenicity in the white matter.

These generated topics are in agreement with the basic themes and topics that were observed in the text reports during the manual classification stage. When the number of topics was increased, we observed that many topics were either redundant or had subtle differences that we could not tease apart with the associated keywords. When the number of topics was decreased, the topics appeared to be a forced consolidation of multiple topics (e.g., hemorrhaging and ventriculomegaly).

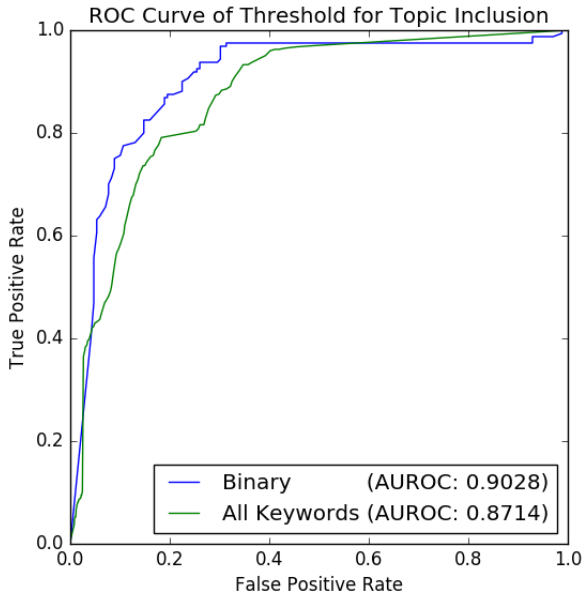

Fig. 3

Quantitatively, the threshold value ϕ was modulated from 0 to 1 to generate ROC curves, plotted in Fig. 3, with the true positive and false positive rates on the y and x-axis, respectively. The blue curve corresponds to the single-label classification task of normal vs. not normal, while the green curve corresponds to the multi-label task. The LDA model does significantly better than the no-discrimination line (not plotted, $45°$ line from the origin) for both tasks, indicating that the model is better than random guessing. The AUROC of the single-label and multi-label classification tasks were 0.90 and 0.87, respectively. When fixing the threshold at $\phi = 0.25$, we observed that the LDA model obtained an accuracy of 84% and 74% accuracy on the single-label and multi-label tasks, respectively.

Though this approach was reasonably successful, a potential pitfall lies in that we have imposed a set of labels onto the topics, i.e., selected a mapping of h. One difficulty with LDA is that the latent topics may not correspond to distinct topics that we as humans may expect. Not only is there the chance that we have applied incorrect labels in our h mapping, it may be that the topics do not correspond to a human-identifiable topic to begin with, undermining the process. We found good alignment between our notions of latent themes in the corpus with most of the topics found by LDA, but there were other confusing ones, such as Topic 3, which had the keywords "normal", "extra-axial fluid", "ventriculomegali". Another point of concern is that the labeling throughout this study was performed by non-experts. Nevertheless, we are encouraged by these results, especially considering Problem 5 of Homework 2, in which we learned that binary classification with partially mislabeled data could still be performed with reasonable accuracy, given enough samples.

Another interesting observation that we did not consider beforehand is that because LDA uses a bag-of-words model, it

may associate a document with a particular topic regardless of a negation in syntax. For example, a report containing the text "no_hemorrhaging or ventriculomegaly" will be assigned to topics associated with both hemorrhaging and ventriculomegaly, even though the report states that these were not observed. To achieve our goal of automatic labeling of text reports, a future approach will require better NLP to take such syntactic intricacies into account.

VI. Conclusion and Future Work

In this work, a small set of manually labeled reports and images was used to train a deep CNN to classify images as "normal" and "not normal". For the AlexNet and VGG16 CNNs, test images were classified with 76% accuracy, making these networks better than the baseline value of 65%. This indicates that the network learned some important features for abnormality detection. In addition to using more of the image set for training and training for more epochs, more work can be done to improve classification accuracy of the neural network. Online repositories contain the training weights of many common CNNs on the CIFAR image set. Literature has shown that these weights can be used for the initialization of other training sets to decrease training time and lead to more accurate networks. If the architecture used in this research can be properly matched to those networks, these weights can be used. Additionally, more architectures such as GoogLeNet[12] and ResNet[13] may be explored.

Furthermore, we have developed a framework for automatically labeling physicians' text reports as additional text reports and images are incorporated into the data set. We have additionally used LDA to discover several latent themes and topics that are present throughout the reports, and used these topics to automatically generate labels on text reports with an accuracy of 84% and 74% on classification tasks with the "normal" keyword and with all keywords, respectively, for a threshold of $\phi = 0.25$. Future work will be focused on improving the syntactic analysis and consulting an expert to obtain more precise manual labeling. We also plan to explore the multimodal-RNN image embedding models that are currently being used in image captioning applications.

References

[1] "AIUM Practice Parameter Neurosonography in Neonates and Infants," 2014.

[2] H. C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image Deep Mining on a large-scale radiology database," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1090–1099, 2015.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] D. Demner-Fushman, S. Antani, M. Simpson, and G. R. Thoma, "Design and development of a multimodal biomedical information retrieval system," *Journal of Computing Science and Engineering*, vol. 6, no. 2, pp. 168–177, 2012.

[7] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Proc. NIPS*, pp. 1–9, 2013.

[9] A. Krizhevsky and H. Sutskever, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems:1097-1105*, 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] Y. J. P. S. S. R. D. A. D. E. V. V. A. R. Christian Szegedy, Wei Liu, "Going deeper with convolutions," *Computer Vision and Pattern Recognition*, 2014.

[13] S. R. Kaiming He, Xiangyu Zhang and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, 2015.