RSNA Intracranial Hemorrhage Detection

Abhinav Aggarwal
aabhinav@student.ethz.ch
18-748-079

Ankush Panwar
apanwar@student.ethz.ch
19-763-051

Pratyush Singh
psingh@student.ethz.ch
19-762-988

Vaibhav Krishna
vaibhavkrishna@ethz.ch
15-947-906

***Abstract*—In this report we present a method to correctly predict presence of Intracranial Hemorrhage and identify its type. Our model imitates the procedure followed by radiologists to analyse a 3D CT scan in real-world. The model utilizes multi-window 3D context from neighboring slices to improve predictions at each slice and subsequently, aggregates the slice-level predictions to provide patient diagnosis for Intracranial Hemorrhage . Our proposed architecture performs significantly better than standard single window based non-contextual models.**

***Index Terms*—RSNA, Deep learning, CNN, CT, Intracranial Hemorrhage Detection**

I. Introduction

Intracranial hemorrhage is the bleeding that occurs inside the cranium and is a serious health problem requiring rapid and often intensive medical treatment. Deaths by intracranial hemorrhage exceeds 20,000 people annually in the United States. Each year, it affects approximately 12-15 per 100,000 individuals. Therefore, it is important to precisely and quickly detect the presence of Intracranial hemorrhage for timely treatment of patients. Researchers have leveraged the deep learning methods to accurately detect the presence of hemorrhage. Identifying the location and type of any hemorrhage present is a critical step in treating the patient. Diagnosis requires an urgent procedure. When a patient shows acute neurological symptoms such as severe headache or loss of consciousness, highly trained specialists review CT scans of the patient's cranium to look for the presence, location and type of hemorrhage. The process is complicated and often time consuming. In this project, we build an algorithm to detect acute intracranial hemorrhage and its subtypes. With the completion of this project we are able to help the medical community identify the presence and type of hemorrhage (Intraparenchymal, Intraventricular etc.) in order to quickly and effectively treat affected patients. Our major contributions are (i) Approach to merge important information using different windows highlighting brain, blood and bone structure of CT scans. (ii) Usage of CNN+sequence model to learn 3D context from neighboring slices.

II. Literature Overview

In recent years, deep learning infrastructures for automatic Intracranial hemorrhage (ICH) detection have based ICH prediction upon either the entire 3D Head CT volume [1] or each 2D CT slice [2]. 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. Intracranial hemorrhage image attenuation significantly overlap with those of gray matter, meaning that simple thresholding is ineffective.

Since direct processing of 3D vortex can require a lot of resources, hence combination of CNN and LSTM are becoming popular in medical computer vision community. The combination of CNN architectures with LSTM has been extensively explored for tasks that require modeling long-term spatial dependencies within image e.g. image captioning [3] and action recognition [4]; or temporal dependencies between consecutive frames e.g. video recognition tasks [5]. Recently, a few studies on biomedical imaging have explored this architecture for leveraging inter-slice dependencies in 3D images [6] [7]. Our approach is more similar to the latter approaches in that it models 3D aspect of medical images as a sequence.

Most of the research has been done for ICH detection using only brain window [8] [9] [10]. In this work we have explored blood and bone windows as well, which has led to significant improvement in predictive performance of model

III. Models and Methods

We introduce our idea and its detailed implementation in this section. We first explain the dataset used for Intracranial Hemorrhage Detection task in Section III-A. We then introduce data preprocessing step in Section III-B. The model architecture and training procedure is detailed in Section III-C.

A. Dataset

The rich image dataset is provided by the Radiological Society of North America (RSNA®) in collaboration with members of the American Society of Neuroradiology and MD.ai. Four research institutions provided large volumes of de-identified CT studies that were assembled to create the dataset: Stanford University, Thomas Jefferson University, Unity Health Toronto and Universidade Federal de São Paulo (UNIFESP), The American Society of Neuroradiology (ASNR) organized a cadre of more than 60 volunteers to label over 25,000 exams for the dataset.

The training data is provided as a set of image Ids and multiple labels, one for each of five sub-types of hemorrhage, plus an additional label for *any*, which should always be true if any of the sub-type labels is true. There are 6 rows per image Id. The label indicated by a particular row looks like *[Image*

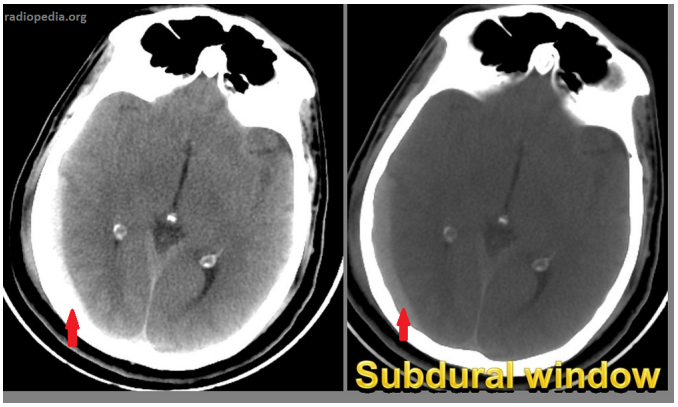


Fig. 1. Importance of correct windowing to detect hemorrhage (Arrow marking subdural hemorrhage). Here left image is taken using standard brain window and right image is created using subdural window

Id]_[Sub-type Name]. All provided images are in DICOM format. DICOM images contain associated metadata. This includes *PatientID*, *StudyInstanceUID*, *SeriesInstanceUID* and other features.

B. Data Preprocessing

Since raw data is present in DICOM (dcm) files, so some preprocessing needs to be done to prepare it for training. Initial step is to extract image and relevant metadata from dcm file. Here we provide little bit of background information about DICOM format to better understand the information retrieval and image extraction process.

Digital Imaging and Communications in Medicine (DICOM) - an international standard related to the exchange, storage and communication of digital medical images. It stores 16 bit images with values ranging from -32768 to 32767 and these values have direct correlation with Hounsfield Scale. Apart from image data it has other valuable parameters such as "Window Center", "Window Width", "Rescale Intercept" and "Rescale Slope". These parameters can be used to compute window intervals which can then be used to highlight specific type of tissue. Generally radiologists have look at five different window intervals for each CT scan/slice.

- Brain Matter window : $W : 80$, $L : 40$
- Blood/subdural window: $W : 130 - 300$, $L : 50 - 100$
- Soft tissue window: $W : 350$–400, $L : 20$–60
- Bone window: $W : 2800$, $L : 600$
- Grey-white differentiation window: $W : 8$, $L : 32$ or $W : 40$, $L : 40$

Think of a window as an instruction to the computer to highlight only voxels which fulfill a specific value. L = window level or center and W = window width or range. For example for brain matter window voxel, display range is between 0 and 80 (Lower limit $= 40 - (80/2)$, upper limit $= 40 + (80/2)$)

As shown in fig 1, choice of correct window interval can be very crucial for hemorrhage detection. Taking inspiration from the fact that correct windowing can be important for ICH detection, we used above mentioned calculations to get

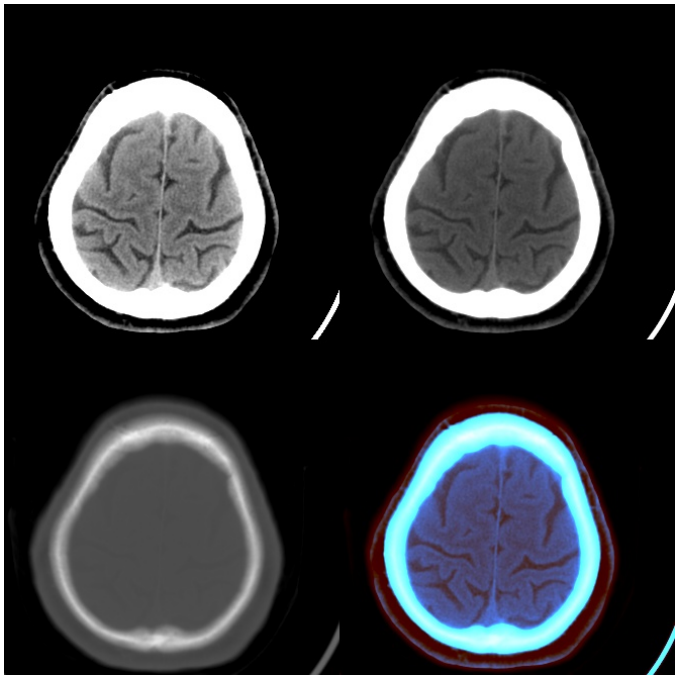


Fig. 2. Top Left: Brain matter window, Top Right: Subdural/Blood Window, Bottom Left: Bone Window and Bottom Right: Final Combined Image used for training

gray scale image for Brain Matter window, Blood/subdural window & Bone window and concatenated them channel wise to convert them into 3 channel image which can then be fed into our model.We selected these three channels as they are considered most critical for ICH detection. After image concatenation we resized the image to 224×224 due to memory constraints and faster training. Fig 2 show images for each of the three mentioned windows and final concatenated images which is being used for training. Finally, extracted images, their image ids and labels are stored in sharded tfrecord files.

Lastly, we also extracted metadata from dcm files and key features like PatientID, ImagePosition and SOPInstanceUID (same as image id) are stored in separate csv file. This extracted metadata will then later be used for sequence model which we will talk in Section III-C2.

C. Training

As shown in fig 3, our architecture is divided into two parts. First part is base model, which is trained using some pretrained architecture and second part is sequence model, whose input is the embeddings learned from base model.

1) Base Model: For base model, we have used pretrained ResNet-50 [11] and Inception-ResNet [12] model. Further we randomly perturbed the small proportion of input images with various geometric transformations like Horizontal Flip, Scaling, Shifting, Rotation and Transposition. This helps in making the model more robust towards geometric changes and also improves predictive performance. This model is trained

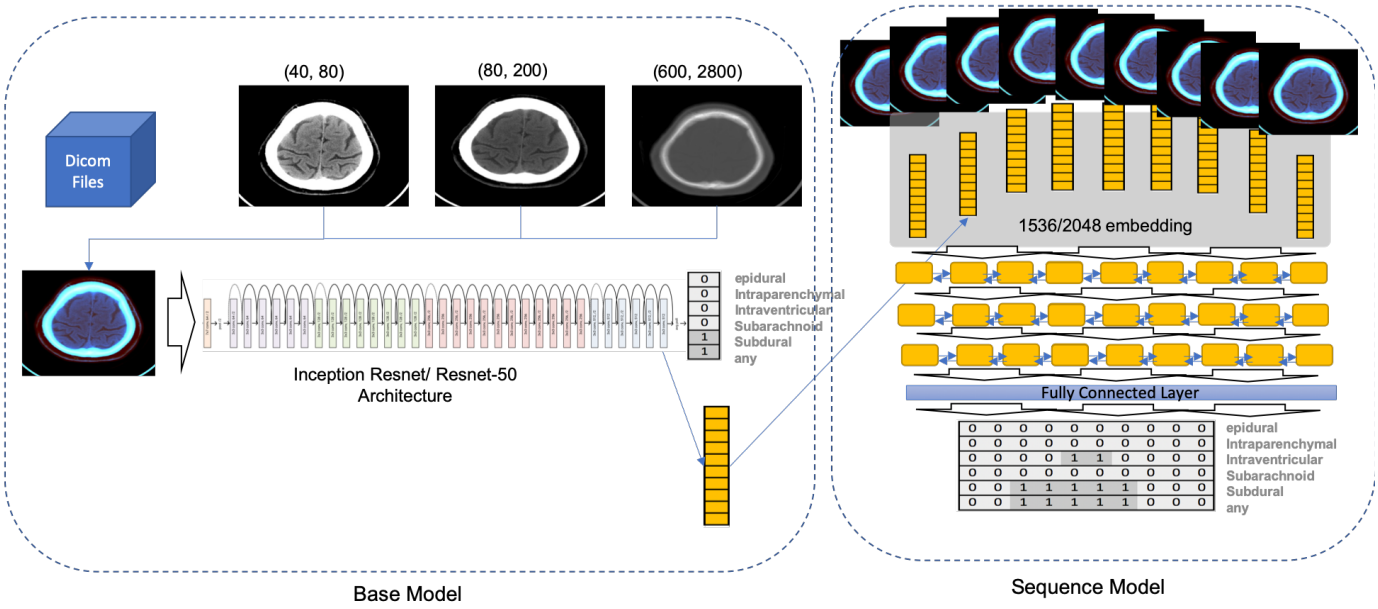


Fig. 3. Schema of the our architecture. As shown in figure, raw dicom files are processed to extract three grayscale images corresponding to brain, blood and bone. These images are then concatenated to create 3 channel image which is then fed to base model (Inception-ResNet or ResNet-50). Then last fully connected layer just before prediction layer is used as embedding for sequence model. Finally these embeddings are fed into multi-layer Bi-LSTM model for final prediction. Note that for simplicity the sequence model architecture in this figure only has nine unrolled states for Bi-LSTM while the actual architecture has sixty states

PatientID	SOPInstanceUID	ImagePosition	Labels
ID_0027f0b7	ID_e1eddd1e7	176	[0,0,0,0,0,0]
ID_0027f0b7	ID_b8d1d1dca	181	[0,0,0,0,0,0]
ID_0027f0b7	ID_b83dea08d	186	[0,0,0,0,0,0]
ID_0027f0b7	ID_3dc7c6292	191	[0,0,0,0,0,0]
ID_0027f0b7	**ID_9b0ba4d90**	**196**	**[1,0,0,1,0,0]**
ID_0027f0b7	**ID_1099fcc1c**	**201**	**[1,0,0,1,0,0]**
ID_0027f0b7	**ID_2ac3ba6e8**	**206**	**[1,0,0,1,0,0]**
ID_0027f0b7	**ID_cbaf2d62d**	**211**	**[1,0,0,1,0,0]**
ID_0027f0b7	ID_d6fa25468	216	[0,0,0,0,0,0]
ID_0027f0b7	ID_846aef7fe	221	[0,0,0,0,0,0]
ID_0027f0b7	ID_33c0c94d1	226	[0,0,0,0,0,0]
ID_0027f0b7	ID_ee2de3062	231	[0,0,0,0,0,0]

TABLE I
METADATA FOR ONE PATIENT

for 7 epochs with learning rate of 0.0001. We used binary cross entropy loss for all the six labels per image and optimized it with Adam optimizer. Training took around 20 hours on a single Nvidia GTX 1080Ti GPU.

Main goal of this task is to extract embeddings for each CT scan/slice which can then be fed to sequence model. We have also evaluated the trained base model to highlight the importance of getting context from neighboring slices. Results can be seen in table II

2) Sequence Model: The Sequence model architecture models the inter-slice dependencies between 2D slices of CT scans of each patient by incorporating bidirectional LSTM [13] layer. The output of the last fully connected layer in base model architecture is passed through multi-layer bidirectional LSTM layer before sending to fully connected layer for final prediction. In this way, the network models the 3D context of CT scan in a small neighborhood, while predicting for single slice. Detailed analysis of metadata for each PatientID and the work done by Grewal et al [8] inspired us to use this approach. Table I clearly demonstrates the importance of sequence model as neighboring images generally have same labels, so learning inter-slice dependencies using neighboring information can further improve the prediction accuracy. This methods also imitates the process used by real world radiologists where various CT scans are generated when patient enters and get out of the machine. Then all these CT scans are analysed to accurately identify whether patient has Intracranial hemorrhage or not.

To create sequenced data for LSTM, we used metadata extracted from dcm files. *ImagePosition* is being used to get temporal information for each patient. Since different patients can have different number of slices, hence we used padding for patients who have number of slice count less than model time steps (60). In case slice count is more than time steps, then regular interval sampling is done, so that temporal structure can be preserved. Other hyperparameters like number of layers, cellsize, dropout rate are selected using ablation studies mentioned in Section VI. This model is trained for 10 epochs with learning rate of 0.00001 with Adam optimizer. This training took around 1.5 hours on single GPU. Hyperparameters used are: Number of Layers: 2, Cell Size: 256 and Dropout Rate: 0.3.

Model	Validation	Test
Inception-ResNet (1-window)	0.106	0.119
Inception-ResNet+Sequence (1-window)	0.080	0.094
ResNet-50 (1-window)	0.123	0.135
Inception-ResNet+Augmentations (3-window)	0.075	0.071
Inception-ResNet+Sequence+No Augmentations	0.052	0.067
Inception-ResNet+Sequence+Augmentations	**0.050**	**0.065**

TABLE II
AVERAGE LOG-LOSS FOR EACH TYPE OF INTRACRANIAL HEMORRHAGE. LOWER SCORE IS BETTER

Model	F1-Score
Inception-ResNet (1-window)	-
Inception-ResNet+Sequence (1-window)	0.913
ResNet-50 (3-window)	-
Inception-ResNet (3-window)	-
Inception-ResNet+Sequence+No Augmentations	0.935
Inception-ResNet+Sequence+Augmentations	**0.942**

TABLE III
F1 SCORE FOR PATIENT LEVEL HEMORRHAGE PREDICTION

IV. RESULTS

In this section, we present results and comparison with other baseline models.

We have tested our model performance by submitting our test predictions on kaggle competition from where this dataset it being taken and reported its private leaderboard score. It evaluates average log loss for each predicted probability versus its true label. Moreover, to evaluate model's performance in predicting whether a patient has hemorrhage or not, we have computed F1-score at patient level. Since we do not have true labels for test data, hence F1-score is reported only on validation set. We used similar evaluation criterion as RADNET paper [8] to predict whether patient has hemorrhage or not i.e. a 3D CT scan was declared positive if the deep learning model predicted hemorrhage in consecutive three or more slices. Also, since F1-score is reported at patient level and not at slice level, hence it is only reported for sequence models.

V. DISCUSSION

Results are presented in Table II and Table III. Inception-ResNet+Sequence+Augmentations models performs best both on log-loss and F1-score on validation and test data. Three window approach sees significant improvement in model's prediction performance as compared to single window approach. Also since base model based on Inception-ResNet performs better than ResNet-50, hence we consider only Inception-ResNet for other experiments.

Addition of Bi-LSTM based Sequence model on top of base model further improves the performance but one interesting point is that it leads to significant reduction in validation loss but test loss does not change that much. This suggests that model might be overfitting due to increased number of parameters and improvement in prediction accuracy is not due to getting information from neighboring slices. This aspect needs to be analysed further to assess the importance of sequence model on top of base model. Further data augmentation during training using random geometric transformations leads to slight improvement and makes model robust.

Our method is two stage approach where first base model is trained, then embeddings are extracted from this trained model. Then these embeddings are used sequence model. One interesting approach can be end-to-end training of both base model and sequence model.

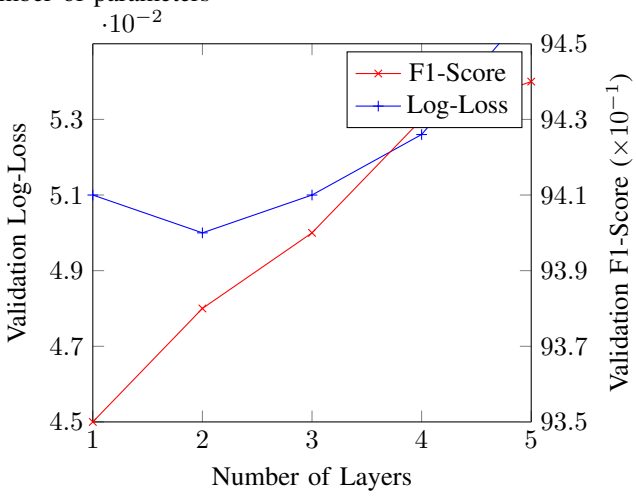
VI. ABLATION STUDIES

In this section, we perform ablation experiments over a number of hyper-parameters for sequence model to select the best possible combination. We have trained all the models for 5 epochs as model performance does not improve much after this. Also each hyperparameter combination is run 5 times and average results are reported to get robust values. Results are shown in bi-axial plot where left y-axis is log-loss and right y-axis is F1-score for validation data.

A. Effect of Number of Bi-LSTM Layers

In this section , we explore the effect of number of Bi-LSTM layers have on model performance. All other hyperparameters remains same i.e cell size = 128 and dropout rate = 0.5.

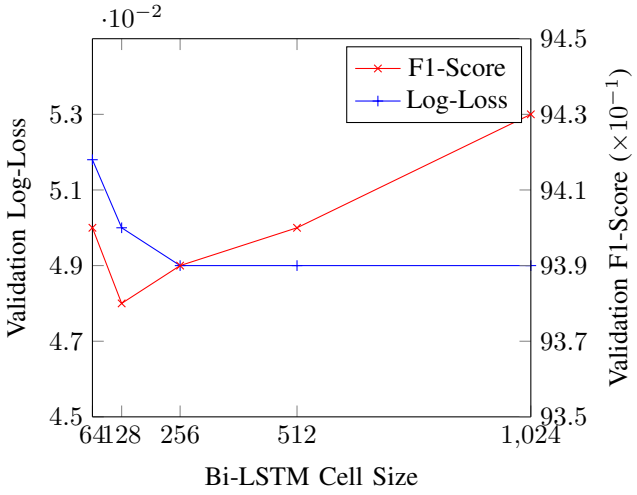
We can see that there is not much effect of layer count on model performance. One surprising thing is that both log-loss and F1-score are increasing with layer count. Here **layer count 2** gives best combination of good prediction performance and number of parameters



B. Effect of Bi-LSTM Cell Size

In this section , we explore the effect of Bi-LSTM cell size have on model performance. All other hyperparameters remains same i.e layer count = 2 and dropout rate = 0.5.

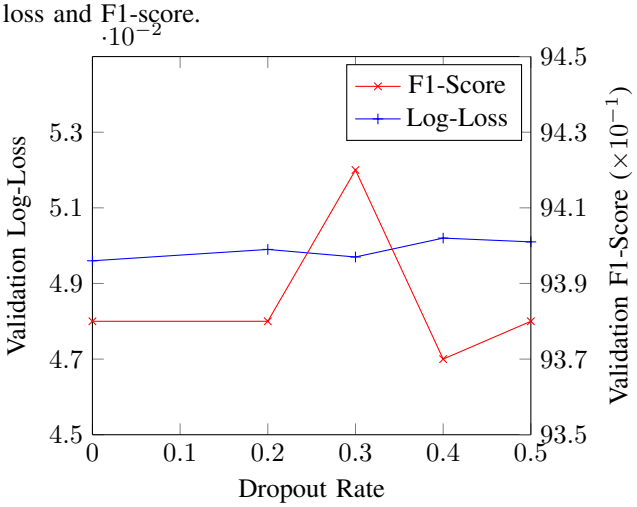
We can see that there is not much effect of change in ceel size in model performance but still **cell size 256** gives best tradeoff between good prediction accuracy and number of parameters



C. *Effect of Dropout rate*

In this section, we explore the effect of dropout rate on model performance for validation data. All other hyperparameters remains same i.e layer count = 2 and cell size = 128.

As we can see there is not much difference with change in dropout rate but still **0.3 dropout rate** gives best combination of loss and F1-score.



VII. Conclusion

In this report we presented a new architecture to combine multi-window and temporal information and introduced a two stage pipeline to solve the problem of Intracranial Hemorrhage Detection. We have shown that these methods lead to an improvement of the prediction accuracy at both CT scan level and patient level compared to simple baseline models. In particular multi-window approach has resulted in significant improvement over standard single window approach.

In future, further analysis of contribution of sequence model needs to be done as it is not clear whether the improvement is due to more number of parameters or getting more contextual information from neighboring slices. Since the dataset has only labels for hemorrhage types, we believe training for segmentation task using our multi-window approach can further improve model's predictive power. Furthermore, currently our architecture is two stage process but end to end training of both base model and sequence model can improve robustness and prediction accuracy

Acknowledgements

We would like to thank the Radiological Society of North America (RSNA) and Kaggle for providing such a huge dataset which can help a lot in taking forward research in this field.

References

[1] K. Jnawali, M. R. Arbabshirani, N. Rao, and A. A. Patel, "Deep 3D convolution neural network for CT brain hemorrhage classification," in , ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 10575, Feb 2018, p. 105751C.

[2] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Development and validation of deep learning algorithms for detection of critical findings in head ct scans," 2018.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CoRR*, vol. abs/1411.4555, 2014. [Online]. Available: http://arxiv.org/abs/1411.4555

[4] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," *CoRR*, vol. abs/1704.07595, 2017. [Online]. Available: http://arxiv.org/abs/1704.07595

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *CoRR*, vol. abs/1411.4389, 2014. [Online]. Available: http://arxiv.org/abs/1411.4389

[6] P.-P. Ypsilantis and G. Montana, "Recurrent Convolutional Networks for Pulmonary Nodule Detection in CT Imaging," *arXiv e-prints*, p. arXiv:1609.09143, Sep 2016.

[7] J. Chen, L. Yang, Y. Zhang, M. S. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation," *CoRR*, vol. abs/1609.01006, 2016. [Online]. Available: http://arxiv.org/abs/1609.01006

[8] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan, "RADNET: radiologist level accuracy using deep learning for HEMORRHAGE detection in CT scans," *CoRR*, vol. abs/1710.04934, 2017. [Online]. Available: http://arxiv.org/abs/1710.04934

[9] L. Prevedello, B. Erdal, J. Ryu, K. Little, M. Demirer, S. Qian, and R. White, "Automated critical test findings identification and online notification system using artificial intelligence in imaging," *Radiology*, vol. 285, p. 162664, 07 2017.

[10] M. Arbabshirani, B. Fornwalt, G. Mongelluzzo, J. Suever, B. Geise, A. Patel, and G. Moore, "Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration," *npj Digital Medicine*, vol. 1, 12 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *CoRR*, vol. abs/1603.05027, 2016. [Online]. Available: http://arxiv.org/abs/1603.05027

[12] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[13] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1422–1432. [Online]. Available: https://www.aclweb.org/anthology/D15-1167