Multi-modal Transformer Learning Medical Visual Representation - An Application Study in Detecting Underrepresented Cardiopulmonary Conditions

Qingxi Meng, Bin (Claire) Zhang
Department of Electrical Engineering
Stanford University, CA
{qingxi,zhangbin}@stanford.edu

Jean Benoit Delbrouck[*]
Department of Biomedical Data Science
Stanford University, CA
{jeanbenoit.delbrouck}@stanford.edu

Abstract

Medical imaging has a challenging problem called hidden stratification, referring to the subsets within a medical task that are visually and clinically distinct and have inherently different implications for treatment. Hidden stratification is difficult for deep learning techniques to tackle as available datasets only have coarsely defined classes and insufficient labeled data for minor classes. Existing work commonly relies on either manually creating annotations for medical variants, which can be laborious, or applying transfer learning from large datesets like ImageNet, which may not work well on medical images due to their specialized nature. In our work, we propose to tackle the hidden stratification problem using the critical information in both medical images and medical text to learn the medical visual representation and detect underrepresented conditions. More specifically, we implemented a multimodal transformer architecture and tested its ability to identify rare cardiopulmonary conditions. We trained and tested our model with MIMIC-CXR dataset to conduct binary classification on whether there is "No Finding" in the sample. Our model achieved a high validation accuracy of 94.16%, surpassing other baseline models. Our further analysis revealed that having both text and image features for the multimodal model is critical for good performance, and increasing the dataset improves the model performance.

1 Introduction

The application of artificial intelligence (AI) in medical imaging diagnostics has shown promising efficacy and efficiency [1]. With current computer vision algorithms and large datasets, automated tasks such as disease detection and classification are made possible and achieving great performances. For example, deep learning has been used for detecting Alzheimer's disease in brain MRI data [2]. However, there is a challenge called the hidden stratification, meaning that the data contains unrecognized subsets of cases [3]. For example, a lung tumour can be solid or sub-solid. Such variations are often visually distinct on the medical images, represent rare and often high-risk medical cases. Hidden stratification is very challenging for AI to tackle due to the severe lack of corresponding labeled data and the extreme difficulty for researchers to label all variants. Therefore, it leads to the interesting question of how to innovate AI to detect visual variants which represent uncommon disease states.

We propose to tackle the hidden stratification challenge in medical imaging by utilizing the rich information in both medical images and text, as well as the relationship between the two. Such approach was hardly considered in past methods as they heavily focused on the medical images

[*] This author is not enrolled in the class

exclusively. Traditional methods of this nature include creating large manually annotated datasets with comprehensive medical variants, transfer learning from large datasets such as ImageNet, and contrastive representation learning.

We propose to use a multimodal transformer to learn the relationship between images and text and to classify the labels of interest, starting from just one label. Our expected contribution is to provide a new method to detect rare cases in medical images without manual annotation and to test whether medical text can provide useful insight to medical imaging tasks. More specifically, we look at whether the model can detect underrepresented cardiopulmonary conditions from radiopgraphs and radioreports, starting from one label of whether there is medical finding in the data. In terms of technicality, this work could bring us insights into how image-related tasks can be improved by incorporating other forms of data. In terms of medical significance, this work would allow us to explore the potential of state-of-art models on difficult disease classification and detection tasks.

2 Related work

2.1 Manual Annotation for Medical Variants

Datasets have been manually created to generate more labels to cover various medical variants. For example, Shih et al. [4] created a huge dataset of expert annotations on the chest radiographs. Similarly, Wang [5] built an open access benchmark dataset of 13,975 CXR images across 13,870 patient cases with experts' annotations. Directly using these labeled datasets would also be not ideal for us as the number of samples in those datasets is too small for a model as complex as ours. Also manually annotating images is time-consuming and require significant help from medical professionals, a procedure that could be very laborious.

2.2 Transfer Learning

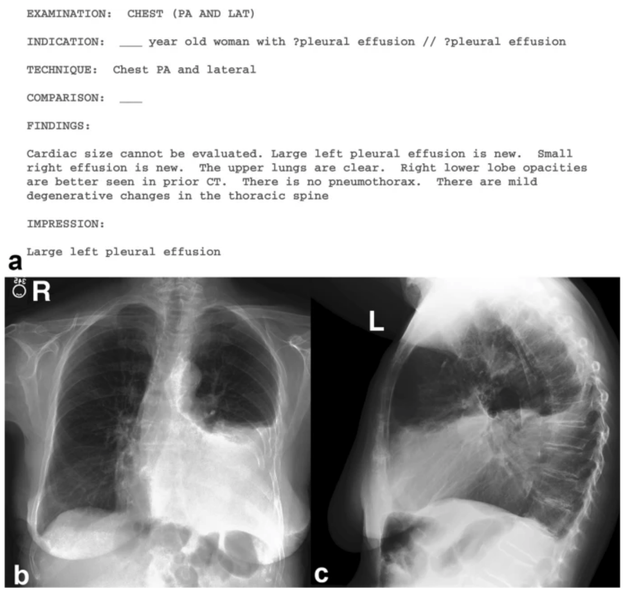
Transfer learning is often used as a workaround to the small datasets. For example, Esteva et al. [6] pretrained the convolutional neural network (CNN) with the ImageNet datasets and then fine tune the CNN for classification of skin cancer. Similarly, Wang et al. [7] also used transfer learning from ImageNet for classification and localization of common thorax disease. However, features in medical images could be very different from those in daily-seen images so it could be suboptimal for us to use a model pretrained on data without complex medical features.

2.3 Self-Supervised Learning

Other work try to use contrastive representation learning models pretrained from natural images. The pretrained models are then used to learn visual features in medical images (e.g. visual variants). For example, Bootstrap Your Own Latent (BYOL) [8] uses self-supervised image representation learning and trains the neural network to predict the target network representation of the same image under a different augmented view. SimCLR (a simple framework for contrastive learning of visual representations) [9] uses a contrastive learning framework to learn useful visual representations. Similarly, MoCo (Momentum Contrast) [10] builds a more complicated version of contrastive learning for unsupervised visual representation learning. Although self-supervised learning works around the problem of lacking labeled data for minor classes, most commonly used models still only utilize the information in images. It would be worthwhile to try a different paradigm that can utilize the insights from other forms of medical data such as text.

2.4 Combine Image and Text Information

Another approach is to use the textual reports together with the medical images, which we will implement for our task, to take advantage of the rich information in textual data besides the visual data. VisualBert [11] is a framework for modeling vision-and-language tasks. However, VisualBert is not specific to medical images, so we will adapt the model to fit our specific task. There are also other work that use mulitimodal modeling for biomedical applications. For example, Hsu et al. [12] focused on unsupervised multimodal representation learning across medical images and reports. This work is similar to ours and uses the same dataset, but we will focus on the supervised learning techniques. Similarly, Wang et al. [13] used the radiographs and radioreports of chest X-rays together to learn specific features for lung disease and relied on recurrent network to extract text embeddings. Similarly, we will utilize both textual and visiual data. But differently, we choose to use Bert [14] model because Bert enables the pretrained deep bidirectional representations and outperforms recurrent network on many sentence-level and token-level tasks.



```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

Figure 1: One sample of MIMIC-CXR radiograph and radioreport.

Dataset	Training	Validation	Test
Number of images	369188	2732	5239
Frontal	248285 (67.3%)	1759 (64.4%)	3708 (70.8%)
Lateral	120756 (32.7%)	972 (35.6%)	1524 (29.1%)
Other	147 (0.0%)	1 (0.0%)	7 (0.1%)
Number of reports	222952	1596	3326
with a finding	149150 (66.9%)	1004 (62.9%)	2753 (82.8%)
Number of patients	64588	500	295
with a finding	38470 (46.9%)	293 (46.0%)	286 (60.6%)

Figure 2: Summary of the downloaded MIMIC-CXR data, which are split into training, validation, and test sets.

3 Data

We use the MIMIC-CXR, a large publicly available dataset of chest radiographs with free-text radiology reports [15].The source of the data is the Beth Israel Deaconess Medical Center Emergency Department. MIMIC-CXR contains a total of 227,835 imaging studies for 65,379 patients between 2011-2016. Each imaging study contains one or more images, usually a frontal and a lateral view, resulting in a total of 377,110 images. The specific statistics of MIMIC-CXR dataset is shown in Figure 2. Each image has a corresponding free-text radiology reports that describe the radiological findings provided by experienced radiologists. Each image also has 14 binary labels suggesting whether the image contains a particular cardiopulmonary condition. One example of MIMIC-CXR radiograph and radioreport is shown in Figure 1. We will use the radiographs to extract image features, and use the radioreports to extract textual embeddings.

4 Approach

4.1 VisualBert

The specific multimodal transformer model we implemented is called VisualBert (Figure 3). VisualBert utilizes the core idea of a transformer's self-attention mechanism to implicitly find correlation between input text and images. Such strength of VisualBert becomes especially beneficial as we want to use the rich information from both medical images and text, as well as their relationship, to identify the challenging hidden stratifications from the radiographs. The original VisualBert takes images and text as inputs, and applies encoders to obtain test embeddings and image features, which are then fed into the attention layers. In our implementation, however, the encoders were not included. Images and text from MIMIC-CXR are encoded to features and embeddings using a separate ResNet-50 (implemented by another group). Then these image features and text embeddings are concatenated and directly fed to the attention layers of VisualBert. On the output end, VisualBert uses a few average pooling and dense layers to conduct the classification. As we are conducting binary classification on one label, the output of the model is either 0 or 1, indicating whether there is or is not findings in the input radiographs and radioreports which is one of the 14 labels in MIMIC-CXR.

We refactored and modified Facebook Research MMF's implementation of VisualBert in PyTorch [16] into a module taking in text embeddings and image features. We did not apply the pre-training steps as described in the original VisualBert implementation for simplicity.

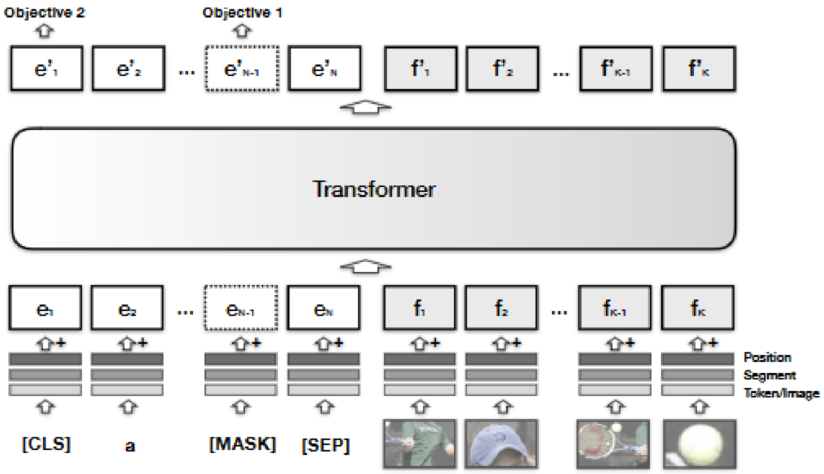


Figure 3: Architecture of VisualBert. Text and image region embeddings are fed into the stack of transformer layers to allow self-attention to find the relationship between vision and language.

4.2 Training Pipeline

As shown in Figure 4, the whole training pipeline of our model consists of 4 different parts. We feed the MIMIC-CXR radiographs and radioreports to our training pipeline. The image feature extractor is a ResNet model pretrained with the MIMIC-CXR dataset once. The pretrained image feature extractor will extract the high-level features from the input radiographs. The Bert embedder is a pretrained embedding generator that we directly use from the Bert model. We have the freedom to choose from the several provided Bert's pretrained embedding configurations. As explained more in the experiments section, we eventually choose to use the Bert's uncase large embedder after some experiments. The input of the Bert's embedder is a sequence of words. The outputs of the Bert's embedder are token IDs and input mask. The token ID of an word specifies the index of that word in Bert. The input mask of a word specifies whether or not the attention layer in the model should pay attention to that word. It is worth to note that we fix the length of our embedding to be 128 because the longest sequence in our dataset has a length smaller than 128. Sequence with words smaller than 128 will be zero padded to this length.

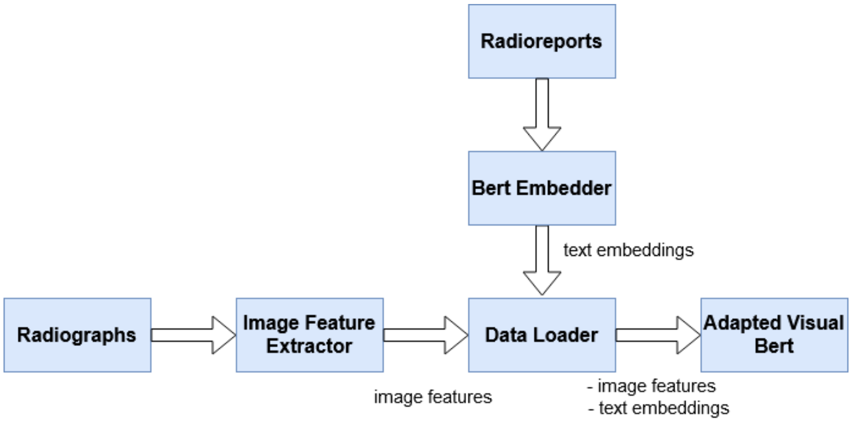


Figure 4: The flowchart of our training pipeline

After the image features and text embeddings are generated, we will feed that into our dataloader. The dataloader will also extract the labels from the radioreports. In our case of binary classification, the dataloader will generate a label ’0’ for data with ’no finding’ under the ’study’ column and generate a label ’1’ otherwise. Finally, the dataloader will be used by our model for training.

5 Experiments

5.1 Hyperparameter Tuning

To first train and tune the model to achieve a validation accuracy as high as possible, seven hyperparameters were tuned: data size, learning rate, batch size, optimizer, loss, Bert configuration and number of hidden layers. For the learning rate, we started with 10^{-5} and gradually increased it by multiplying a factor of 10 each time. The best learning rate observed for 100,000 datapoints was 10^{-2}. A learning rate scheduler that reduced the learning rate on plateau was also trialed, but did not work well in our case as accuracy was not improving as well as a fixed learning rate. Four different batch sizes were trialed: 4, 8, 16 and 32. Only relatively small batch sizes were trialed due to compute constraint on Google Colab. Although the small batch may be noisy, it could also provide regularizing effect[19]. In the end, a batch size of 16 was chosen for a good trade off between the stability and memory usage. Four optimizers were trialed for our model. Adam, Adamx, and RMSprop optimizers did not show promising results. Eventually SGD optimizer was found to work the best for our model after fine tuning. For the loss function, cross entropy loss was implemented as to fit our binary classification task.

5.1.1 Model Configuration

It is critical to choose a good Bert configuration for the Bert part within the VisualBert model. The Bert’s pretrained small uncased configuration did not work well since there were many complex medical terms in our texts. After switching to Bert’s pretrained large uncased configuration, we observed a clear improvement in validation accuracy trend. After trying some other more complicated pretrained Bert configurations, we found out the general Bert configuration worked the best for our text data. For the number of hidden layers, we started with 1 and gradually increased it to 100. However, with increasing number of hidden layers, the computational cost also increased tremendously but not too much accuracy increase was observed. Thus, 2 hidden layers were used in the end, which worked well on the model.

5.1.2 Best Model

We used the best hyperparameters we found and trained our model with 100,000 datapoints and 90 epoches. The best validation accuracy we achieved was 94.16%. Detailed training processes were shown in Figure 5 and Figure 6. Figure 5 shows that the training loss was fluctuating while decreasing, potentially due to the randomness in the stochastic descent implemented. Another potential reason could be that those were not enough data for this complex model. As shown in the Figure 6, the validation accuracy smoothly increased as the number of epoches increased. We stopped training at epoch 90 because it reached a plateau and the validation accuracy stayed nearly the same.

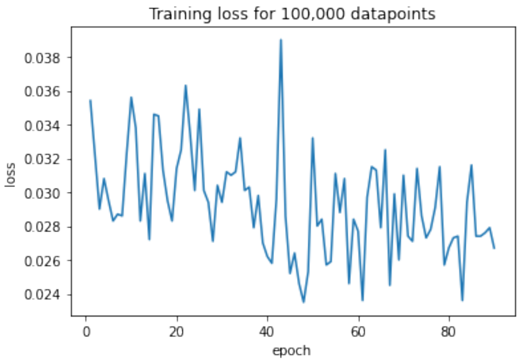


Figure 5: Training Loss for 100,000 Data

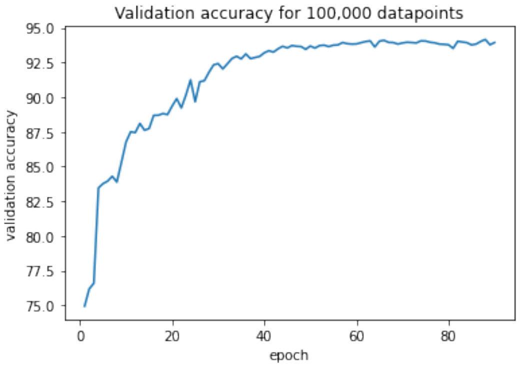


Figure 6: Validation Accuracy for 100,000 Data

5.2 Impact of Data Size

Data size here refers to the number of image features generated as input to the proposed model. Two different data sizes were trialed: 30,000 and 100,000. Here We used the same parameters as mentioned above for both models and trained them for 50 epoches. The results are shown in Figure 7 and Figure 8.

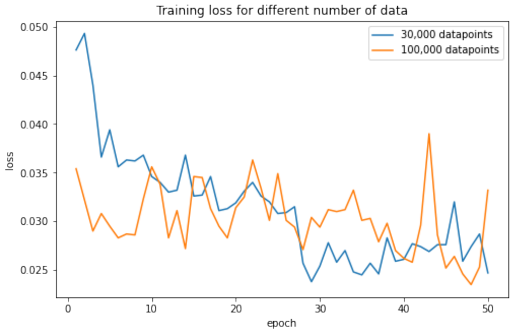


Figure 7: Training Loss for Model Trained with Different Data Sizes

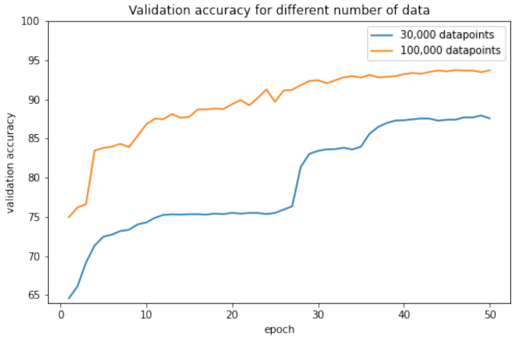


Figure 8: Validation Accuracy for Model Trained with Different Data Sizes

As shown in Figure 7, the loss curves for models trained with both data sizes gradually decreased. However, the loss for the model with larger data size was fluctuating more than the other. The reason may be that it was harder to train and stabilize the model with much more datapoints. In Figure 8, the validation accuracy of the model trained with 100,000 datapoints was higher than that of the 30,000 through the whole training process, suggesting that the model benefited significantly from having more datapoints.

5.3 Baseline Comparison

Our proposed model was compared with two different baselines. Since the input image features for our model were extracted with a ResNet-50, we chose the ResNet-50 as a baseline. Specifically, we fed the extracted image features directly into a ResNet for the classification task. This baseline did not use the text data. As shown in Table 1, the baseline with ResNet-50 has the best validation accuracy around 78.81%, worse than the best accuracy of our model, suggesting that the proposed model may have benefited from the text embeddings and its complex structure. More details about the loss and accuracy of the Resnet-50 baseline are shown in Figure 9 and Figure 10.

Table 1: Comparison with Baselines

Part		
Name	Description	Best validation accuracy
Our model	Adapted VisualBert	∼94.16
Baseline 1	ResNet-50	∼78.81
Baseline 2	Average Pooling + Dense Layer	∼76.28

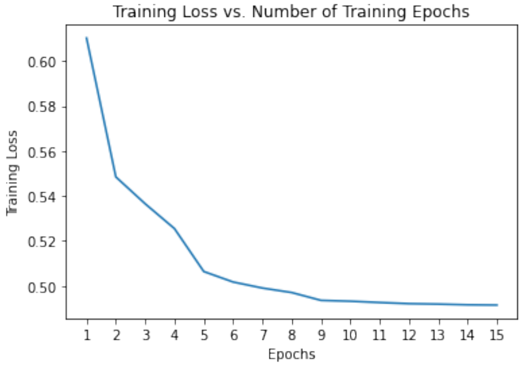


Figure 9: Training Loss for baseline 1

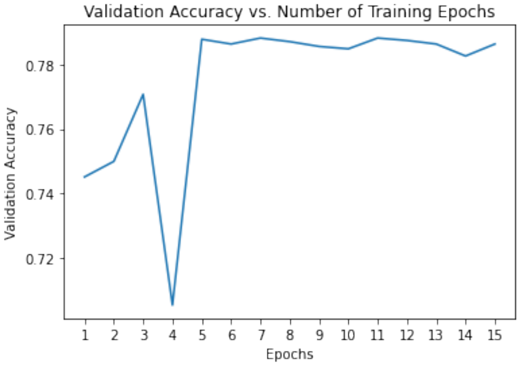


Figure 10: Validation Accuracy for baseline 1

The second baseline we compared with is an extremely simplified version of our model. Since the Bert part is the core component in our mode, it would be meaningful to compare with a model without the Bert part. We created such a baseline with one one average pooling layer and one dense layer. The parameters for both two layers were the same as the those used in our model. The details about the loss and accuracy of this baseline are shown in Figure 11 and Figure 12. We only trained it with 20 epoches because this baseline quickly reached the plateau. As shown in Table 1, the best accuracy for this baseline is around 76.28%, which is worse than that of our model. This result also suggests that the attention layers in our model are crucial for achieving good performance.

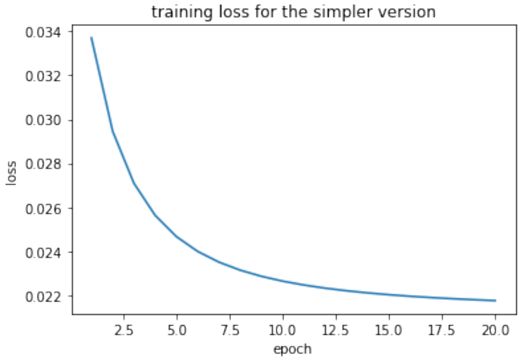


Figure 11: Training Loss for baseline 2

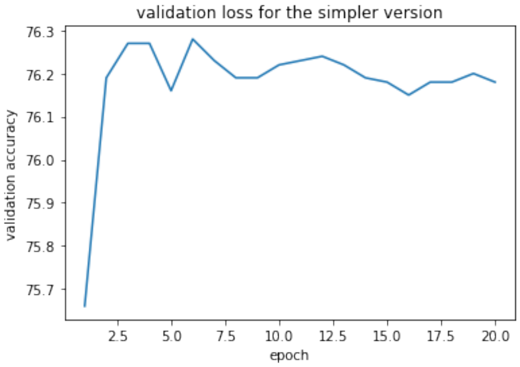


Figure 12: Validation Accuracy for baseline 2

5.4 Ablation Study

Since our multimodal model takes inputs from both image features and text embeddings, it would provide us insights to the model by testing how much each kind of input contributes to the model performance. Therefore, we performed two experiments. For the first experiment, only the image features were fed into our model. Similarly, for the second experiment, only the text embeddings were fed into our model. The training results for both experiments are shown in Figure 13 and Figure 14. From Figure 13, we can see that the training loss for input with either image features only or text embeddings only is stagnant. Also, from Figure 14, the validation accuracy for input with either image features only or text embeddings only is also stagnant. These results show that the proposed model needs both the image features and the text embeddings to work well.

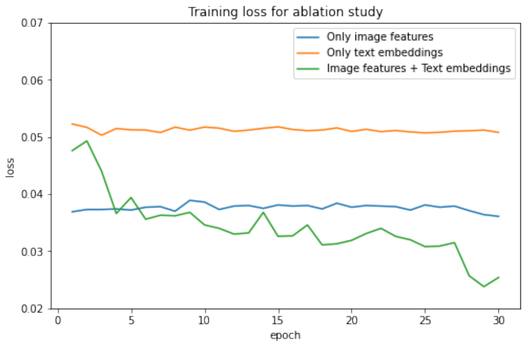


Figure 13: Training Loss for ablation study

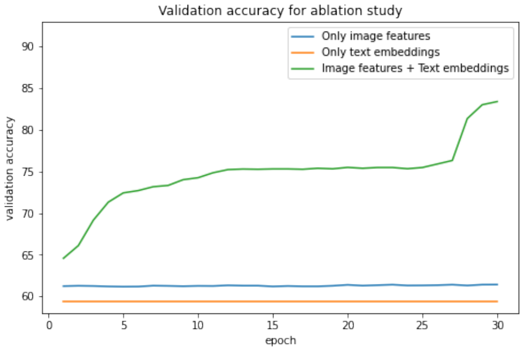


Figure 14: Validation Accuracy for ablation study

6 Conclusion

We presented a multimodal transformer for learning medical visual presentations from pair of radiographs and radioreports. We created our model by adapting from an existing implementation called VisualBert. For the binary classification task, specifically the binary classification of whether MIMIC-CXR radiographs and radioreports contain cardiopulmonary condition findings, our model outperforms two baseline methods and achieves a high accuracy around 94.16%. To further improve the proposed model's performance in the future, we could train our model with datapoints far more than the maximum amount (100,000) we trained with, if compute allows. In addition, other metrics including confusion matrix, sensitivity, specificity, and F1 score of our model performance should be investigated in the future.

Another interesting future work would be to train our model for multi-label tasks. For example, we could test our model on classification of all 14 labels in the MIMC-CXR dataset. Results from this experiment can be compared with those ranked on the leaderboard of CheXpert [17], a large dataset of chest X-rays as well as a competition of automated chest X-ray interpretation for the same 14 cardiopulmonary conditions. Several of the highest ranked models include SuperCNN Ensemble, Hierarchical-Learning-V1 ensemble [18], Conditional-Training-LSR ensemble, etc, which all achieved an average AUC as high as 0.929 or above. We could run our model on the CheXpert benchmark and gain more insights into the capability of the proposed model.

7 Contribution

Claire, Qingxi and Jean-Benoit were highly involved with the discussion for our design. Jean-Benoit provided us with dataset we used throughout the project. Claire, Qingxi and Jean-Benoit extracted VisualBert model and implemented the dataloader and the training pipeline. Claire and Qingxi worked on training the model with different parameters. Claire and Qingxi wrote all parts of this paper.

References

[1] Pesapane, Filippo et al. “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine.” European radiology experimental vol. 2,1 35. 24 Oct. 2018, doi:10.1186/s41747-018-0061-6

[2] Islam J, Zhang Y A novel deep learning based multi-class classification method for Alzheimer’s disease detection using brain MRI data. In: International Conference on Brain Informatics, 2017. Springer, pp 213-222

[3] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL ’20). Association for Computing Machinery, New York, NY, USA, 151–159.

[4] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence, 1(1):e180041, (2019).

[5] Linda Wang and Alexander Wong. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. arXiv preprint arXiv:2003.09871, 2020.

[6] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, 2017.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (ICML), 2020a.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[11] Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

[12] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. arXiv preprint arXiv:1811.08615, 2018.

[13] Wang et al. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR, 2018.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[15] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6, 317 (2019).

[16] Singh, Amanpreet, Goswami, Vedanuj, Natarajan, Vivek, Jiang, Yu, Chen, Xinlei, Shah, Meet, Rohrbach, Marcus, Batra, Dhruv and Parikh, Devi, MMF: A multimodal framework for vision and language research,https://github.com/facebookresearch/mmf

[17] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[18] Pham, Hieu H., et al. "Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels." arXiv preprint arXiv:1911.06475 (2019).

[19] Wilson, D. Randall, and Tony R. Martinez. "The general inefficiency of batch training for gradient descent learning." Neural networks 16.10 (2003): 1429-1451.