Multimodal Semantic Embeddings to Reduce Hidden Stratification in Medical Imaging Data

Michael Cooper*
Department of Computer Science
Stanford University
Stanford, CA 94309
coopermj@stanford.edu

Kent Vainio*
Department of Computer Science
Stanford University
Stanford, CA 94309
kentv@stanford.edu

Abstract

The problem of hidden stratification presents a bottleneck for the adoption of machine learning systems within the medical image analysis space: the reliable mis-identification of certain subsets (such as minority classes) by machine learning models in medical imaging reduces trust in such models in an environment where the cost of failure is high. We hypothesize that unimodal data is more susceptible to hidden stratification and that multimodal models, especially classifiers trained on multimodal data, will obtain superior performance at identifying minority classes in medical imaging datasets. Unfortunately, additional data modalities, such as text annotations, are only available after a clinician has evaluated the image, effectively nullifying the scalability potential of predictive learning systems in such environments. To approach this problem, we present MEEMS (Multimodal Embedding and Expression of Medical Semantics), a generative model architecture which allows for the projection of unimodal medical image data into multimodal semantic hyperspace, with the ultimate aim of improving medical image diagnosis through tackling the hidden stratification problem. Our contribution is twofold: first, our architecture provides a novel general-purpose embedding framework to elevate unimodal data into the multimodal domain, and second, we show that such an embedding framework extracts semantically rich information which is not adequately captured by benchmark convolutional architectures. Through comparing the performance of a MEEMS-embedding classifier against a ResNet-50 baseline image classifier, our results indicate that MEEMS embeddings outperform the baseline in binary and 14-label classification. Furthermore, an unsupervised analysis of the embedding space shows that our embeddings do cluster into groups corresponding to the majority labels of the dataset, uncovering some of its stratification. Our visualization of attention weights further adds a greater degree of interpretability to our architecture, which is important in a field where a low tolerance for opacity and high cost of mistakes limits the viability of deploying traditional machine learning image evaluation models at scale. However, despite the incremental performance gains achieved on classification tasks and a relatively clear partitioning of embedding space, our MEEMS embeddings are not able to capture the semantic space of minority classes as accurately as we had hoped, suggesting that future work is needed to tackle the problem of hidden data stratification.

1 Introduction

Correcting the problem of hidden stratification in medical data stands to substantially improve rate of large-scale adoption of machine learning diagnostic models to assist components of clinical care, such as patient triage. Hidden stratification - the tendency of a predictive model to fail on a subsets of data, such as minority classes - is a critical drawback of existing models working to diagnose conditions which are both rare and lethal.

*Equal contribution

BIODS 220: AI + Healthcare (Autumn 2020).

Though much of the prior work in machine learning for clinical diagnostic support relies on traditional deep neural architectures [1], [2], [3], in 2020 Oakden-Rayner et al. demonstrated that such models often suffer from the the problem of hidden stratification when applied to medical data [4]. Given the recency of the Oakden-Rayner hypothesis, there exists an extremely limited body of work oriented around counteracting the hidden stratification problem in statistical diagnosis models.

Our approach builds off recent work by Meng et al. [5] which shows that medical diagnostic models which leverage multimodal data (such as images and paired radiological report data) obtain superior performance compared with those based off unimodal medical data. Though Meng et al.'s work requires textual report data as inputs, our architecture contains a generative captioner, which allows for unimodal medical images to be embedded in multimodal space. We hypothesize that the spatial distribution of learned embeddings in high-dimensional space lends itself well to hyperplanar separability (or near-separability) between class clusters corresponding to output labels.

This is an interesting contribution in the context of the existing landscape both because hidden stratification stands to severely bottleneck the adoption of statistical diagnostic methods in the medical space, and because publicly available datasets are increasingly multimodal. Additionally, visualizing elements of our architecture - such as attention weights, and the spatial distribution of embedding outputs - yields a strong degree of interpretability. This is important in medicine, as the high cost of failure typically means there is a low tolerance for opaque models [6] : our contribution of interpretable metrics, therefore, aligns with the paradigms of interpretability necessary for a model to be deployed at scale in medicine.

Our results demonstrate that a multimodal embeddings-based approach makes some headway in solving the hidden stratification problem, although not as much as we had anticipated. Our architecture yields superior classification performance to baseline methods on binary and 14-label classification (though, in the 14-label case, obtaining superior accuracy on some labels comes at the expense of accuracy of others), and clustering of the embeddings suggests a semantically meaningful partitioning of findings into the "No Finding" and "Finding" group, with the "Finding" group further divided into distinct classes. These results suggest that our embeddings-based approach successfully extracts hitherto unexploited semantic information from input medical images. However, our MEEMS embeddings also fail to fully capture the semantic space of minority classes, which is a critical objective in solving the hidden stratification problem, suggesting that future work is needed in this space.

2 Related Work

There exists a moderate extent of prior work in the domain of multimodal medical diagnosis prediction. Zhi et al. leverage a multimodal attention network which incorporates both ICD (International Classification of Diseases) and procedure codes, as well as textual clinical notes, to assist with diagnostic prediction [7]. Experiments demonstrated that the multimodal setup performed superior to medical unimodal and multimodal comparison baselines, including Dipole [8], Retain [9], DoctorAI [10], PacRNN [11], and a standard multimodal RNN. Further work incorporates both image and textual data to produce state-of-the-art results in embedding generation and classification. Hsu. et al designed a system capable of creating joint representations of medical image and text data in both a supervised and unsupervised manner [12]. Their architecture uses a pre-trained DenseNet to encode images, and various NLP featurizers such as Glove word embeddings to form features vectors out of input text. These features can then be combined using supervised or unsupervised learning techniques. The authors note that even a small degree of supervision can be very beneficial to learning joint representations, which makes us think that learning embeddings in the context of a classification task would be optimal. Wang et al. propose TieNet [13], a novel architecture for extracting text and image representations. This multi-level CNN-RNN architecture with attention is capable of learning meaningful representations between images and text to both generate reports (used in the task of auto-annotation), and classify images, which in this case were thoracic X-rays. This is one of the architectures that heavily influenced our design decisions, as it can effectively project unimodal data into multimodal space.

More recently Zhang et al. developed a system called ConVIRT (contrastive visual representation learning from text), which is a framework for learning visual representations through the naturally occurring relationship between images and textual data [14]. ConVIRT is an unsupervised, domain-agnostic strategy that uses a bidirectional contrastive objective to train. It can be used to pre-train image encoders, leading to strong performance on a range of image-based classification tasks. ConVIRT represents a step forward in the domain of representing images using multimodal information, and is a more light-weight solution to multimodal embedding generation compared to Wang et al's work.

Most recently, in 2020 Meng et al. [5] leverage Li et al.'s VisualBERT architecture [15] to perform multimodal binary classification of medical conditions on the MIMIC-CXR dataset [16], and obtain a 94.2% classification accuracy with their approach. VisualBERT incorporates the core BERT architecture developed by Devlin et al. [17], which employs multiple transformer layers with self-attention to discover complex semantic patterns in provided language input. VisualBERT modifies the input layer by adding a set of visual embeddings to model an image. In Li et al.'s paper, both visual and textual embeddings are passed into the multi-layer BERT architecture, allowing the model to build up a joint representation of multimodal data. Using VisualBERT for the task of generating multimodal embeddings would be ideal, however given time constraints this quarter we ended up creating a system more similar to TieNet. Another important related work that we drew upon is Audebert et al's multi-modal document classification scheme [18]. This model takes in image data of a document, as well as OCR-extracted textual features to produce a multimodal representation of the artefact at hand. We were inspired by their simple use of fully-connected layers to unify image and linguistics features, and so decided to use this for our own embedding architecture.

As will be discussed further in the Approach section, we intend to use an image captioning model to extract linguistic semantic meaning from the images as they are passed into our system. Recent work in this area has leveraged the concept of attention [19] to enable deep models to focus on different components of the image over the course of composing the caption. Xu et al.'s Show, Attend, and Tell paper [20], the architecture which we employ in our system, combined an encoder-decoder approach with an attention mechanism to achieve state-of-the-art image captioning performance on benchmark datasets.

One area of related work which we intend to leverage in our project is that of word embeddings: though general-purpose word embeddigs like Word2Vec enable admirable results on standard natural language tasks, they have been found to obtain inferior performance on medical datasets, since the generating distribution of medical textual data differs greatly from that of Google News, on which Word2Vec was trained [21, 22]. Therefore, relevant literature to our project includes word embeddings trained on medical corpora, such as BioWordVec [23], and those investigated in a biomedical word embedding survey by Khattak et al. [24].

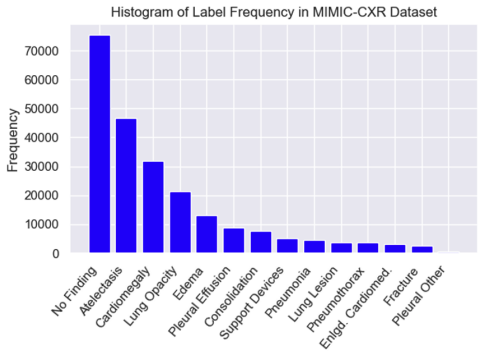
3 Data

For this project, we used the MIMIC-CRX database [16], which is a large, publicly available dataset of 377,110 de-identified DICOM chest radiographs and paired radiological reports. The images are grouped by study, as the data from MIMIC-CRX is sourced from 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. In addition to the radiographs and reports, each study is labelled: each label assigned to an study corresponds with one of the 14 disease classifications shown in in Figure 1. The distribution of these 14 labels suggest a high degree of hidden stratification within the data - there are quite a few minority classes that are important to detect, but which may go unnoticed by traditional AI models. Thus MIMIC-CRX is an ideal dataset for us to test the effectiveness of our multimodal embeddings, especially in the context of uncovering and classifying rarer disease sub-types.

As shown in the histogram in Figure 1, the 14-label annotated version of the dataset is imbalanced: some classes, like "No Finding", "Atelectasis", and "Cardiomegaly" are common, while classes like "Lung Lesion" and "Fracture" are comparatively rare. The binary version of the dataset is also imbalanced: there are 152,372 studies with findings, and 75,455 studies without findings, yielding an (approximate) 2:1 ratio of findings to no-findings among the studies represented in the dataset.

4 Approach

To approach the problem of effectively leveraging multimodal medical data to perform minority-class classification, we devised a generative model architecture to deal with the problem of *retroactive analysis*: given that the paired textual data in MIMIC-III only becomes available after a clinician has examined the image, a purely discriminative model would be incapable of operating in any meaningful predictive capacity, such as aiding with patient triage. The MEEMS architecture therefore consists of three models which collectively perform multimodal medical data classification: (1) A *generative captioner* produces a BioWordVec-embedded [23] radiological report based on an input image; (2) a *multimodal embedder* combines image and generated-report data and produces a corresponding embedding vector; and (3) a *embedding classifier* takes in multimodal embeddings and produces a predicted class label (the final layer is swapped out so that the embedding classifier can perform either binary or multi-class classification). The complete MEEMS architecture diagram is shown in Figure 2.



Processed Report: "no acute intrathoracic process pa and lateral views of the chest provided there is no focal consolidation effusion or pneumothorax the cardiomediastinal silhouette is normal imaged osseous structures are intact no free air below the right hemidiaphragm is seen"

Figure 1: Left: Histogram of lung condition labels in the MIMIC-CXR dataset showing the imbalance in class representation with the 14-label annotations. Right: sample image and associated radiological report drawn from the dataset (the report is processed, meaning that subsections of the report have been merged, and punctuation has been removed).

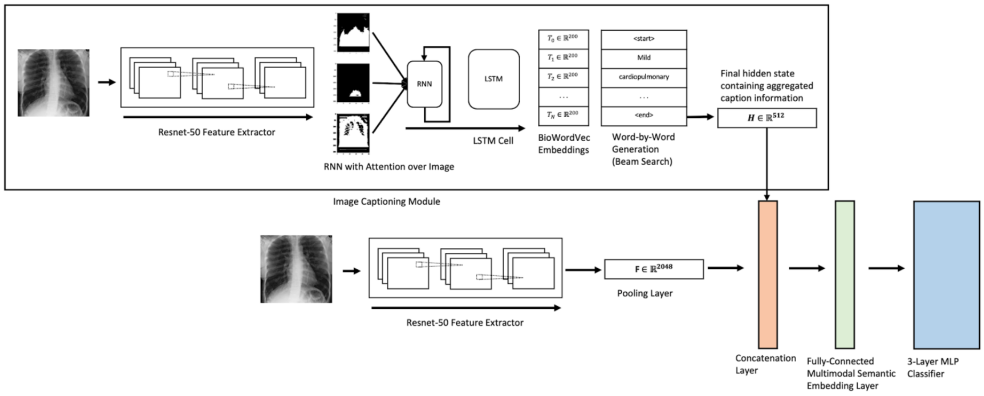


Figure 2: The MEEMS end-to-end embedding and classification architecture.

We partitioned our dataset into the following four subsets: $D^{tr}_{captioning}$ refers to the training set of 103,864 image-report pairs used to train the generative captioner; $D^{tr}_{embedding}$ refers to the training set of 206,888 image-report pairs used to train the baseline models and MEEMS classifier, D^{val} refers to the validation set of 2,813 image-report pairs used to validate the baseline models, generative captioner, and MEEMS classifier, and D^{test} refers to the test set of 4,896 image-report pairs used to test the performance of the baseline models against the MEEMS classifier.

4.1 Generative Captioner

We implemented an image captioning model based off of Xu et al's 2015 Show, Attend, and Tell paper [20, 25], which was selected largely due to its leveraging of visual attention: we hypothesize that visualizing the attention weights used to generate a caption can provide a heuristic on the performance of the captioner (to verify the model places high attention on the heart if detecting enlargement of the cardiomediasteinum), and contribute an element of interpretability to the model. For example, giving a clinician the ability to understand where the model is focusing as it evaluates an image can provide insight into how the model came to its predicted diagnosis; it may also inform other candidate conditions potentially present in the image (a model which focuses frequently on the cardiac region may imply a cardiac abnormality, even if the final predicted diagnosis is not a cardiac one).

Our implementation is composed of a ResNet-50 feature extractor (pre-trained on $D^{tr}_{captioning}$; weights are frozen during captioner training), an attention network consisting of three fully-connected layers (the final fully-connected layer being the *attention layer*), and a decoder consisting of an LSTM sequential layer and four fully-connected layers. Images are first passed through the feature extractor, then the attention network, then the decoder, which

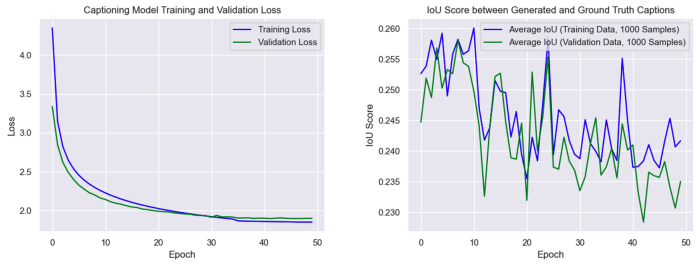


Figure 3: Left: training/validation loss per epoch while training the image captioning module. Right: average IoU score per epoch between generated captions and ground-truth captions from a training/validation sample of 1,000 images. Though the IoU score declines over training, the main takeaway from the righthand plot is that the IoU is a noisy, roughly-constant signal in the context of training the captioning model.

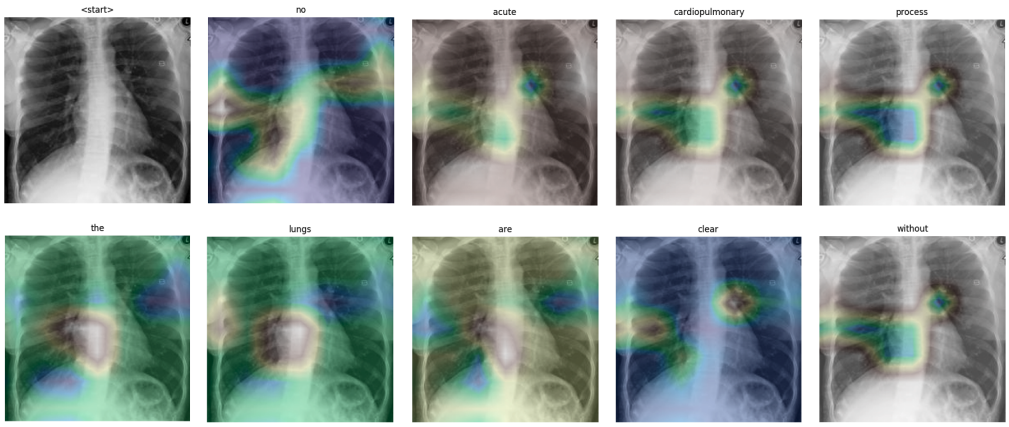


Figure 4: Visualized attention weights corresponding to the first ten words of the generated caption of an image drawn from the training set, along with the generated and ground-truth caption for the image. Observe that there is some semantic correlation between the attention weights and the words being generated: "cardiopulmonary" places high attention on the heart and lungs, and "clear" places high attention on regions of the lung where pneumoniac infilrates are commonly observed [26].

generates a series of scores over the corpus vocabulary (where the corpus consists of all radiological reports in $D^{tr}_{captioning}$).

Once the model has generated scores over the corpus vocabulary, beam search (with a beam size of three) is used to generate a caption from the sequence of scores.

The captioning module was trained (and validated) for 50 epochs on $D^{tr}_{captioning}$ and $D^{val}_{captioning}$. All models in this paper were trained with an Adam optimizer, a learning rate scheduler which reduces the learning rate by a factor of 10 if the validation loss failed to decrease after three epochs (starting at 4×10^{-4}), and early stopping if the validation loss failed to decrease after five epochs. In addition to tracking the loss over each epoch, each epoch we randomly drew 1,000 images from each of the training and validation sets, and calculated the average IoU across the sample between each generated caption and its corresponding ground-truth. This average IoU score was used as a proxy for the accuracy of our generated captions. The plots of train/validation loss over training epochs, and train/validation sample IoU over training epochs, are shown in Figure 3.

4.2 Multimodal Embedder

The multimodal embedder is composed of a concatenation layer, which concatenates the ResNet-50-extracted features from the image with the hidden state output from the captioning module, and two fully-connected embedding layers. After concatenation, we produce a vector $V \in \mathbb{R}^{2560}$. This is then mapped to $\mathbb{R}^{1000}$ through the first fully-connected layer, and finally to $\mathbb{R}^{200}$, resulting in the final MEEMS multimodal semantic embedding.

4.3 Embedding Classifier

During our evaluative experiments, we append a 3-layer MLP classifier to the embedding output of our model. The MLP maps the $\mathbb{R}^{200}$ embedding to $\mathbb{R}^{50}$, and finally to $\mathbb{R}^{out}$. Depending on the evaluative experiment that is being performed, the embedding classifier either produces scores over binary classes $L^{out}_{binary} \in \mathbb{R}^2$, or over a label vector $L^{out}_{multiclass} \in \mathbb{R}^{14}$ encompassing all labels present in the MIMIC-CXR dataset.

5 Experiments

We performed two classes of experiments to test the performance of the MEEMS system. The first set of experiments are *explanatory experiments*, designed to give us more information about the behaviour of our system and the embeddings it produces. The second set of experiments are *evaluative experiments*, designed to measure the performance of MEEMS embeddings against standard image classification baselines at multiple classification granularities.

5.1 Explanatory Experiment 1: Attention Weights Visualization

We visualized the weights from the output layer of the attention network in the image captioning module to interpret how the model was learning to produce captions. In Figure 4, we visualize the attention weights corresponding to the first ten words of a generated caption of an image drawn from the training set (the ground-truth and generated captions are also visualized). The visualization scheme involved calculating the attention weights for each generated word, and mapping the weights to a topographic colour map (blues/purples indicate low attention; browns/whites indicate high attention) which is then resized and overlaid on the image. For this image, we also printed out the generated and ground-truth captions: observe that, though the generated and ground-truth captions are indeed different in phrasing and semantics, there is substantial overlap in the semantic information each conveys (e.g. both note the absence of pleural effusion and the unremarkable countours/silhouette of the mediasteinal region).

5.2 Explanatory Experiment 2: Unsupervised Clustering and Visualization of Embeddings

Next, we embedded the entire training set $D^{tr}_{embedding}$, and, interpreting the embedding vectors as spatial coordinates, visualized them in space. Given each embedding vector $e_i \in \mathbb{R}^{200}$, we normalized the vector and then used Principal Component Analysis [27] to project each embedding vector into both the $\mathbb{R}^2$ and $\mathbb{R}^3$ coordinate spaces. By colouring each embedding in accordance with its label, we sought to understand the extent to which a vector's label was correlated with the spatial location of its embedding. In Figure 5, we showcase these visualizations. In the leftmost sub-image of Figure 5, which represents the 2D embedding space, embeddings with a label of "No Finding" are predominantly found on the left hand side of the spatial mapping; embeddings with a label of "Pleural Effusion" and "Cardiomegaly" are predominantly found in the lower right of the spatial mapping; and embeddings with a label of "Edema" are predominantly found in the upper right of the spatial mapping. In 3D (rightmost sub-image - each has a different camera angle), we once again see a highly distinct "No Finding" region, with certain dominant finding classes, such as "Cardiomegaly" also carving out distinct regions of the coordinate space. In this 3D visualization we also see the "Atelectasis" class forming a very visually distinct subset of points in black. These results correlate highly with the findings from the confusion matrix, which saw many findings being bucketed into the two most common finding categories - namely "Cardiomegaly" and "Atelectasis", both of which form distinct clusters in the PCA visualizations. We also made an animated visualization of the 3D PCA plot to better visualize the embedding distribution over 3D space.

In order to better understand how these groups were clustered in 2D space, we plotted the PCA results for each class separately. The resulting plot can be found in the Supplementary Material, Section 2, since it was too large to include in this report. From this experiment we arrived at a similar conclusion - the "No Finding" label represents a very distinct cluster, and certain finding sub-groups such as those mentioned above are quite prominent. Some

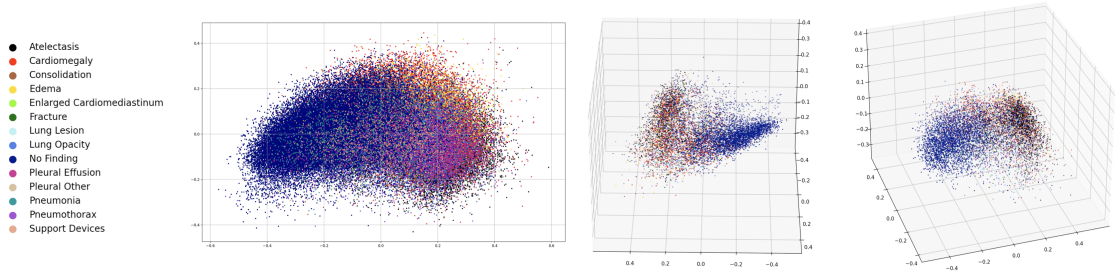


Figure 5: Left: Spatially visualized $PCA_{\mathbb{R}^{200}\rightarrow\mathbb{R}^2}$ reduction of a sample of 1,000 images from the training set, colour-coded according to their ground-truth label. Middle, Right: $PCA_{\mathbb{R}^{200}\rightarrow\mathbb{R}^3}$ reduction of the same data, shown from different angles. Though the clusters are not linearly separable, observe how different labels vary in spatial location within the visualization in both 2D and 3D.

classes, such as "Lung Lesion", or "Pleural Other", which are among the rarest classes, have a widely dispersed distributed clustering, meaning the embeddings did not accurately capture information about these classes.

From the plots, therefore, it is clear that the MEEMS embedding scheme encodes a significant amount of information relevant to the binary label in the embeddings, though no clearly linearly separable clusters are shown aside from perhaps the "No Finding" region. It is possible - though difficult to verify through the PCA visualization - that the spatial distinctiveness of clusters of other category labels is more pronounced in $\mathbb{R}^{200}$, and that the lack thereof in $\mathbb{R}^2$ or $\mathbb{R}^3$ is a function of the dimensionality reduction scheme we employed. Further investigation is needed to determine exactly how well-clustered the embeddings are in their native dimension.

5.3 Evaluative Experiment 1: Binary Classification

As a first evaluative experiment, we compared the performance of a 3-layer MLP binary classifier trained for 50 epochs on MEEMS embeddings of $D^{tr}_{embedding}$ against a ResNet-50 binary classifier trained for 20 epochs on the raw images drawn from $D^{tr}_{embedding}$. Both models were validated on D^{val}, and performance was evaluated on D^{test}. Training/validation loss/accuracy plots can be found in the supplementary materials.

Confusion matrices and ROC curves of the baseline binary classifier and the MEEMS binary classifier are shown in Figure 6. The AUC and F1 scores of each model are shown in the table below:

Model Type	AUC	F1 Score
Baseline	0.76	0.815
MEEMS	**0.79** (micro - **0.89**)	**0.817**

From the results table we see that MEEMS outperformed the baseline in terms of both AUC score and F1 score. The AUC and F1 scores for MEEMS were generated by creating a binarized label matrix, from which AUC and F1 calculations could be performed on a column-wise basis. Note that the MEEMS classifier has two AUC values - one is the average of the "No Finding" and "Finding" AUC values, and the second is the micro-average AUC calculated on a flattened version of the aforementioned binary label matrix.

The ROC curves of the baseline (shown in Figure 6) and MEEMS model show visually similar results, which correlate with the statistics shown in the table above. Analysis of the confusion matrix proves to be more interesting - we observe that the MEEMS classifier - compared to the baseline - obtains a greater true positive rate and false positive rate, and a lesser true negative rate and false negative rate. The confusion matrix is very sensitive to how class thresholds are determined - in this case we take the argmax of the prediction to determine whether a label is a "No Finding" or "Finding", which removes any possibility of applying a carefully calculated threshold that one could obtain from the ROC curve. This leads to the results shown, which may appear to contradict the good AUC score and F1 score that the MEEMS model obtained. However, this confusion matrix is merely calculated at an operating point with high true positive and high false positive rate, which would correspond to the upper-right region of the micro-average ROC curve in dashed-pink.

Even through this operating point is not optimal in all circumstances, assuming medical aid in not in short supply, this is a superior performance in the medical context: given that many of the diseases being classified are severe if untreated - pneumonia, edema, and pleural effusion are all examples - a medical image classification model should optimize for reducing the false negative rate. MEEMS embeddings successfully achieve this objective

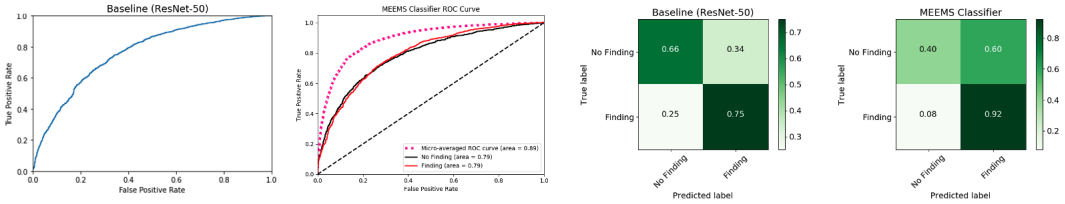


Figure 6: Left: the ROC curves for the ResNet-50 baseline and MEEMS classifier models on the binary classification problem. Right: the confusion matrices for the ResNet-50 baseline and MEEMS classifier models on the binary classification problem.

(MEEMS reduces the false negative rate from 0.25 to 0.08) at the expense of increasing the false positive rate (MEEMS increases the false positive rate from 0.34 to 0.60).

We would also like to note that the ResNet-50 baseline was trained on a larger number of training samples - 370,000 versus the 100,000 used for our MEEMS model. This was done because of time constraints on training during the quarter. However, despite the smaller number of available training examples, we observe that the MEEMS model performs better in terms of the metric we want to maximize (true positive rate).

5.4 Evaluative Experiment 2: 14-Label Classification

As a second evaluative experiment, we compared the performance of a 3-layer MLP 14-label classifier trained for 50 epochs on MEEMS embeddings of $D^{tr}_{embedding}$ against a ResNet-50 14-label classifier trained for 9 epochs on the raw images drawn from $D^{tr}_{embedding}$. Both models were validated on D^{val}, and performance was evaluated on D^{test}. Training/validation loss/accuracy plots can be found in the supplementary materials.

Confusion matrices and ROC curves of the baseline binary classifier and the MEEMS 14-label classifier are shown in Figures 7 and 8. Below we present a table containing the AUC and F1 scores for both models.

Model Type	AUC	F1 Score
Baseline	0.82	0.364
MEEMS	**0.85**	**0.378**

The score table clearly shows that the MEEMS model outperforms the baseline in both AUC and F1 metrics. The AUC, F1 and ROC curves were all calculated using binarized label matrices as described above in section on binary classification. This slight performance edge is confirmed by inspecting the 14-way ROC curves. These were generated by plotting ROC curves for each of the classes in a binarized label matrix. A micro-average ROC was also calculated, using a flattened version of the binarized label matrix. From Figure 7 we see that the MEEMS classifier has better ROC curves for more of the "Finding" classes, at the expense of other less common classes. Notably, we see the model generates a very good curve for "Pneumothorax" (in lilac), however generates a much worse curve for "Enlarged Cardiomediastinum" (very rare class in bright green).

As in the binary case, the confusion matrix presents a narrower picture of model performance, as it uses the argmax function to determine the output class of the model from its predictions, rather than thresholds chosen to achieve a user-defined true positive rate-false positive rate balance. Looking at the 14-way confusion matrix, we observe that both the baseline and MEMES matrices are quite similar - both attribute many labels to "Atelectasis" and "Cardiomegaly", which are the dominant labels in the "Finding" category. Upon close inspection, however, we see that the MEEMS confusion matrix is slightly more diverse in its correlations, and that is has a strong "No Finding" to "No Finding" score.

6 Conclusion

The results of our evaluative experiments demonstrate that a MEEMS-based classifier outperforms a ResNet-50 baseline on both binary and 14-way classification tasks. In addition, our classifier is capable of achieving highly sensitive performance at the cost of increasing the false positive rate relative to the baseline model. In the context of screening against dangerous diseases, this is most certainly a positive. In the 14-way case, our model is prone to classifying images into the majority classes, just like the baseline, however is slightly better at capturing the semantic information. In addition, through PCA analysis we determined that our model can generate semantically

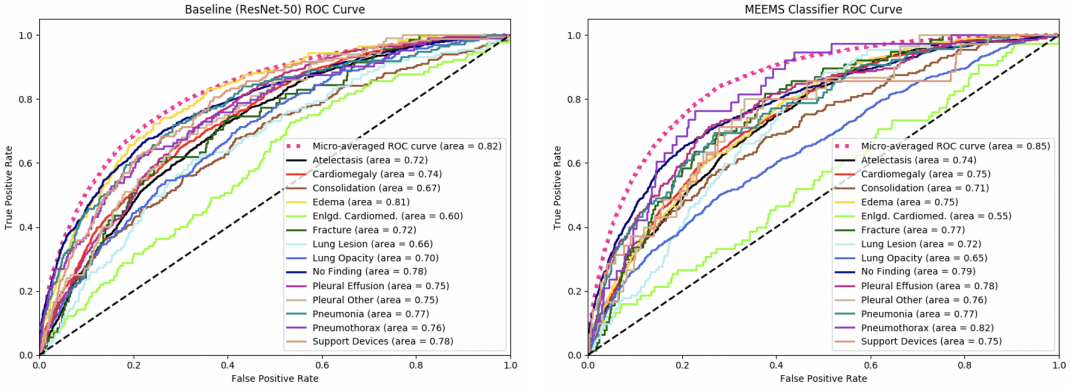


Figure 7: ROC curves for the ResNet-50 baseline and MEEMS classifier models on the 14-label classification problem.

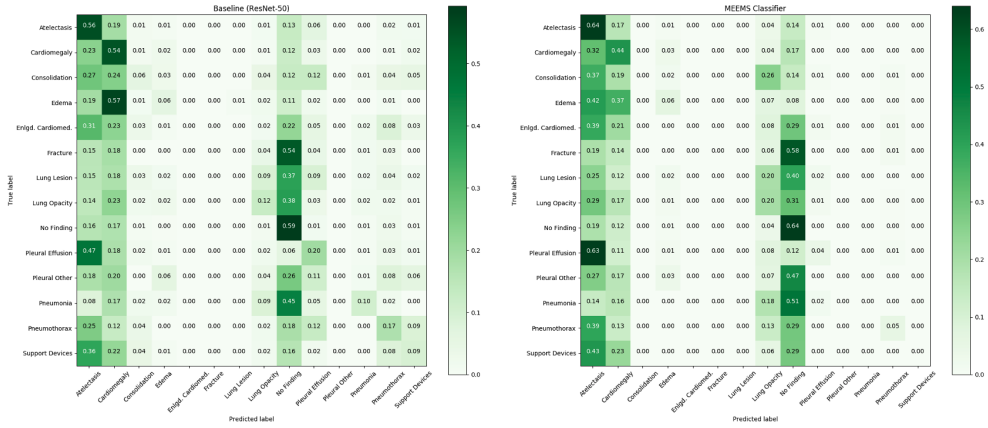


Figure 8: Confusion matrices for the ResNet-50 baseline and MEEMS classifier models on the 14-label classification problem.

meaningful multimodal embeddings whose partitioning in higher dimensional space lies very clearly across binary classification boundaries, and slightly less clearly across 14-way classification boundaries.

Moreover, the results of our interpretive experiments show that our MEEMS architecture is interpretable: attention weights during caption generation can be easily isolated and visualized, and visualizing the embedding space effectively provides high-level intuition into the underlying structure of the dataset on which the model was trained.

Overall our model achieves superior performance on classification tasks than standard baselines, although it does not do as well as we had hoped on minority classes. Therefore, though our approach of generatively projecting unimodal data into multimodal embedding space is a valuable contribution, it does not reduce hidden stratification to the desired extent. Therefore, future work remains in this area.

Two promising avenues for future work include using a transformer architecture (such as VisualBERT) as the Multimodal Embedder within the architecture, and making the generative model symmetric across modalities. Though a modality-symmetric architecture would represent a significantly more difficult problem, such an architecture would, for example, be able to take in a radiological report, from which it would generate an image to pass into the multimodal embedder, and such a model could learn from the significant amount of publically-available textual medical report data which does not always include corresponding images.

7 Contributions

M.C. and K.V. collaboratively ideated the model architecture. K.V. implemented the ResNet-50 feature extractor and the attention network, and integrated the vocabulary with BioWordVec embeddings. M.C. implemented the data loaders, the caption search within the captioning module, and a preliminary LSTM caption classifier (featured in the project midpoint, but not in the final report). Both K.V. and M.C. collaboratively wrote the training loops for the the captioning module, the feature extractor, and the baseline architectures. Both K.V. and M.C. ideated the interpretability visualization metrics: K.V. implemented the embedding visualization module, and M.C. implemented the attention visualization module. Both K.V. and M.C. wrote the final paper.

The authors would like to thank Jean-Benoit Delbrouck, a postdoc in Biomedical Data Science at Stanford University, for providing data access and guidance on this project.

The authors would additionally like to thank Sagar Vinodababu for his Show, Attend, and Tell tutorial on GitHub. This tutorial was referenced extensively in the development of the captioning model in the MEEMS architecture.

References

[1] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[2] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks". In: *Journal of Medical Imaging* 3.3 (2016), p. 034501.

[3] Jens Behrmann et al. "Deep learning for tumor classification in imaging mass spectrometry". In: *Bioinformatics* 34.7 (2018), pp. 1215–1223.

[4] Luke Oakden-Rayner et al. "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 151–159.

[5] Quingxi Meng and Claire Zhang. "Multi-modal Transformer Learning Medical Visual Representation". In: *BIODS 220: AI + Healthcare* (2020).

[6] Xun Jia, Lei Ren, and Jing Cai. "Clinical implementation of AI technologies will require interpretable AI models". In: *Medical physics* 47.1 (2020), pp. 1–4.

[7] Zhi Qiao et al. "Mnn: multimodal attentional neural networks for diagnosis prediction". In: *Extraction* 1 (2019), A1.

[8] Fenglong Ma et al. "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 1903–1911.

[9] Edward Choi et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3504–3512.

[10] Edward Choi et al. "Doctor ai: Predicting clinical events via recurrent neural networks". In: *Machine Learning for Healthcare Conference*. 2016, pp. 301–318.

[11] Zhi Qiao et al. "Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2018.

[12] Tzu-Ming Harry Hsu et al. "Unsupervised multimodal representation learning across medical images and reports". In: *arXiv preprint arXiv:1811.08615* (2018).

[13] Xiaosong Wang et al. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9049–9058.

[14] Yuhao Zhang et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *arXiv preprint arXiv:2010.00747* (2020).

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

[16] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[17] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[18] Nicolas Audebert et al. “Multimodal deep networks for text and image-based document classification”. In: *arXiv preprint arXiv:1907.06370* (2019).

[19] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[20] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. 2015, pp. 2048–2057.

[21] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[22] José Antonio Minarro-Giménez, Oscar Marın-Alonso, and Matthias Samwald. “Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation”. In: *arXiv preprint arXiv:1502.03682* (2015).

[23] Yijia Zhang et al. “BioWordVec, improving biomedical word embeddings with subword information and MeSH”. In: *Scientific data* 6.1 (2019), pp. 1–9.

[24] Faiza Khan Khattak et al. “A survey of word embeddings for clinical text”. In: *Journal of Biomedical Informatics: X* 4 (2019), p. 100057.

[25] Sagar Vinodababu. *A PyTorch Tutorial to Image Captioning.* May 2018. URL: `https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning`.

[26] Parul Pahal and Sandeep Sharma. “Typical Bacterial Pneumonia”. In: *StatPearls [Internet]*. StatPearls Publishing, 2019.

[27] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.