AI in Medicine I

Practical exercise 3

Submitted to the
Department of Informatics
I31 - AI in Medicine and Healthcare
I32 - Computational Imaging in AI and Medicine
of TUM

by
Argudo, Mateo (03717216)
Bolaños, Daniela (03765336)
Frey, Daniel (03632203)
Javadov, Aydin (03749463)
Otto, Julia (03698083)

December 12, 2023

1 Lowdata Regime and Autoencoders

1.1 Lowdata Performance

In order to convert the data multi-label from ChestMNIST, binary classification model was used, in this case, it was interested to have 2 classes: Healthy and diseased. Then, the implementation of the accuracy was done, and the distribution of the data with the new model can be observed in the Figure 1.

According to the Figure, the distribution is performed with equilibrate data, as the difference between both of them differs only in approximately 5%, so, the accuracy obtained for the model could be appropiate accuracy in the binary classification. However, it would be considered as a Class Imbalance since the accuracy perform only 58% of the test model; so, in order to alleviated the bias towards a majority class, i.e., healthy class in the Figure, a modification of the loss function could be stablished to increase sensitivity towards the minority group (i.e., diseased class in the Figure).

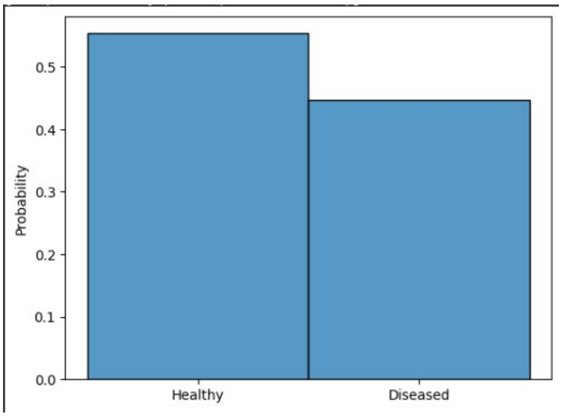
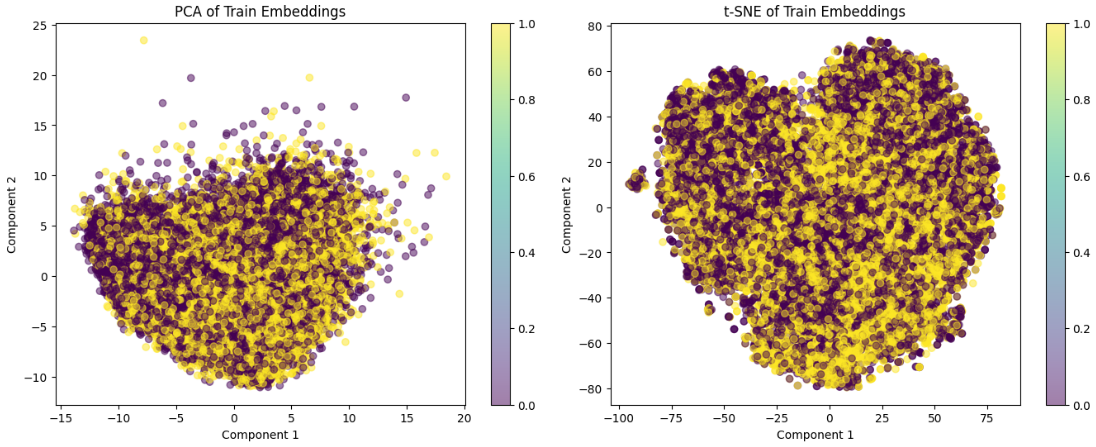


Figure 1: Data distribution of the accuracy of the binary classification model (healthy vs diseased) of ChestMNIST.

The better way to select a model that would allow to set a very high number of epochs without overfitting could be do an *Early-stopping.* This is a regularization/optimization technique used to prevent overfitting and improve the generalization of neural networks; with this technique, one part of the training set is keep as the validation set. Hence, when the performance on the validation set is getting worse, it has to stop inmediately the training on the model, otherwise, it will be prone to overfit to the train set. In addition, it is neccesary to get the minimum value of the loss function, since the idea is to decrease it.

1.2 Examining the Latent Space

Latent space is examined based on the embeddings of batch inputs to the classification network. Two unsupervised dimensionality reduction techniques, namely Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), are applied to these embeddings. As shown in Figure 2, the methods lack to cluster the latent space information with respect to the provided labels. The model was (over)trained on a limited selection of 300 samples, which is too few to allow any meaningful encoding.



(a)

(b)

Figure 2: Latent embeddings of a simply trained (300 samples) classification network after dimensionality reduction (22433 to 2) with PCA (a) and t-SNE (b).

1.3 Autoencoder Performance

Taking a look at the reconstruction from the autoencoder, it can be seen that the shape of the structures match but the image produced is slightly more blurry than the original test image (see Figure 3). This leads to the assumption, that the latent space allows for a good encoding of larger patterns but missing detail in the transitions. The learned weights from the encoder are transferred to the classification model, aiming to provide start values fitting the image space. Results show, that with fine tuning a higher accuracy on the test set can be achieved than with random weight initialization. The accuracy on the training data is slightly lower. This indicates that the model is less prone to overfitting. Whilst fine tuning we experienced that the model is sensitive to different values p.

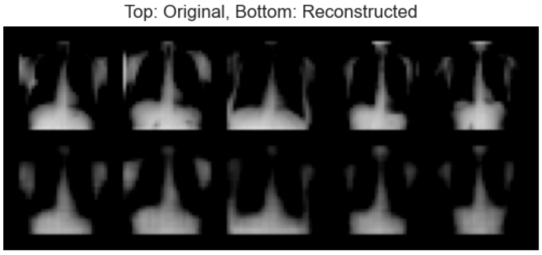


Figure 3: Comparison of original test images on the top and reconstructed images by the autoencoder at the bottom.

2 Transfer Learning and Freezing Strategies

2.1 Freezing Networks

The idea of transfering the weights of the chest trained encoder (*model_ae*) and freezing them in the pneumodia model (*model_pneumonia_transfer_frozen*) excepting the fully connected layer is to test if the encoder part of the autoencoder has captured some important features in its convolutional layers. And most importantly, whether this important features translate not only into chest classification, but also into pneumonia detection. The results seem to confirm the hypothesis that the autoencoder captures relevant general features of the chest that can be transfered to the pneumonia model achieving around 85% accuracy with the weights freezed.

Analysing the latent space we can see how the data is distributed. For PCA. See Fig. 4 its shown how a compact cluster is formed close to 0 for component 1 and for middle to low values of component 2. This ensures that component 1 is more precise in separating pneumonic and healthy patients. tSNE is also used to analyse the latent space. In Fig. 5 it illustrates 2 clear groups and how pneumonia is closely related within itself because of the low distances between neighbors.

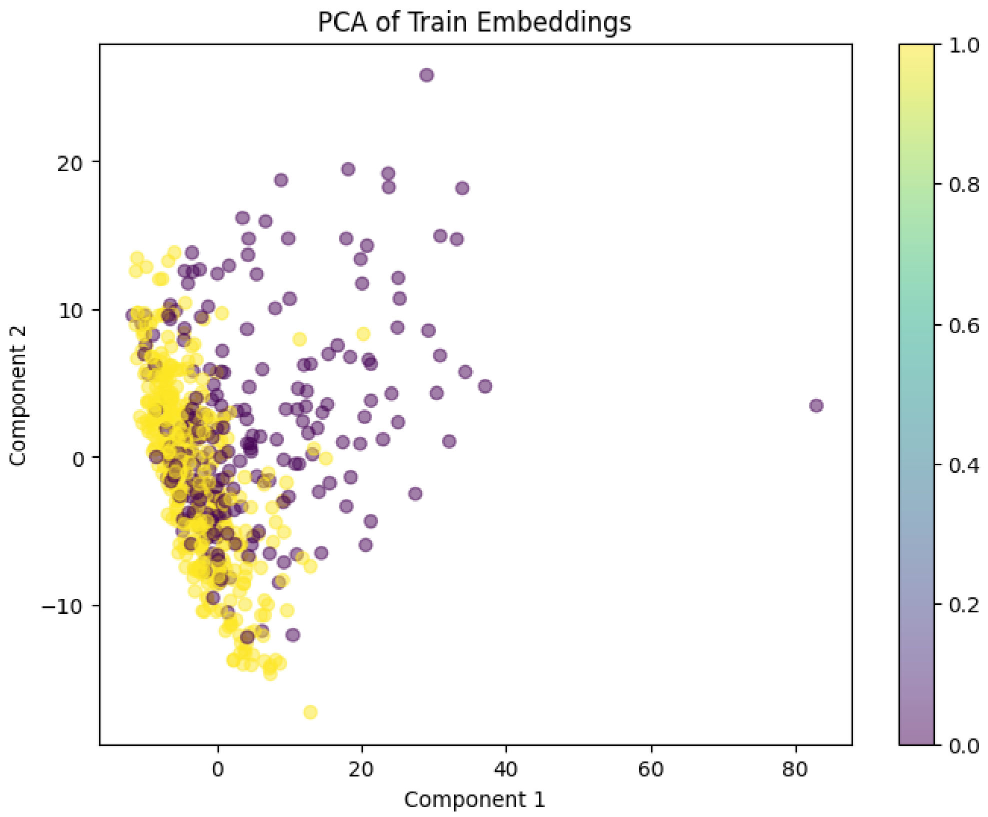


Figure 4: Component one separates healthy from pneumonia patients stronger than Component 2. Although there is mixture in the lower range of component 1. For component 2 high ranges seem to indicate pneumonia.

2.2 Trainable Networks

In this task, we're highlighting the difference between 'transfer learning' and 'fine-tuning.' We do this by unfreezing the frozen parameters from the previous task and keeping the weights transferred from the chest model to the pneumonia model as trainable. The accuracy score of approximately 86% in the previous task indicates that freezing the lower layers effectively preserved fundamental feature extractors, allowing the model to learn new domain-specific features primarily in the top layers.

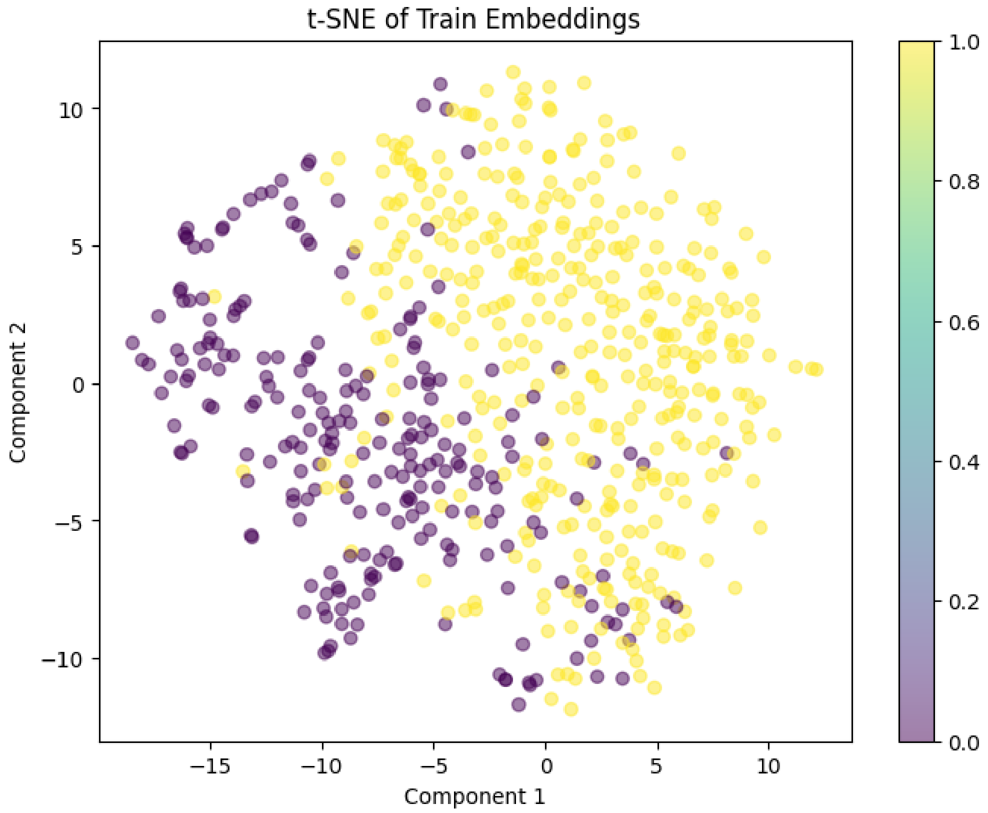


Figure 5: The tSNE shows two clear clusters, with pneumonia data being closely related to each other.

On the other hand, fine-tuning the model by keeping the parameters trainable, rather than freezing them, yielded slightly divergent accuracy scores. In our current codebase, we observe an increased accuracy score of 87%. This shift could be attributed to the model potentially overfitting to the new, smaller domain. However, an intriguing observation surfaced during our development and debugging phase. Upon rerunning the entire process from scratch (without a seed), we noted an accuracy score of 84% for the fine-tuning approach. This discrepancy likely stems from the random nature of the initialization process. What is important is that, we can still provide justification for this behavior. Retraining the entire network, encompassing all layers, introduces the risk of forgetting previously learned features. This potential loss can be ascribed to the comprehensive nature of the retraining process.

Figure 6 and Figure 7 demonstrate the latent space visualizations for the corresponding fine-tuning method.

2.3 Latent Space Musings

We have conducted examination of the latent space using PCA and t-SNE on three distinct occasions. In this report, we compare the outcomes and provide insights into our observations. Notably, the first visualization, which involves the latent embeddings of a modestly trained (300 samples) classification network after dimensionality reduction, reveals a significant difference compared to the second and third visualizations. In this instance, the methods fail to effectively cluster the latent space information according to the provided labels.

The observed deficiency can be attributed to the model being (over)trained on a limited dataset of 300 samples, rendering it insufficient for meaningful encoding. Conversely, when employing transfer learning and fine-tuning methodologies, distinct label clusters become evident in the latent space. Despite similarities—owing to the freezing of all layers except for the fully connected head—the latent representations learned by the chest encoder are largely preserved in the pneumonia encoder. This preservation implies that lower-level features and representations captured by the chest encoder influence the latent space of the pneumonia encoder up to the point of freezing.

While the latent space visualizations exhibit similarities, they are not identical. This discrepancy arises from the pneumonia encoder's capacity to adapt to pneumonia-specific features during training, resulting in a more task-specific and finely tuned latent space for the pneumonia classification task.

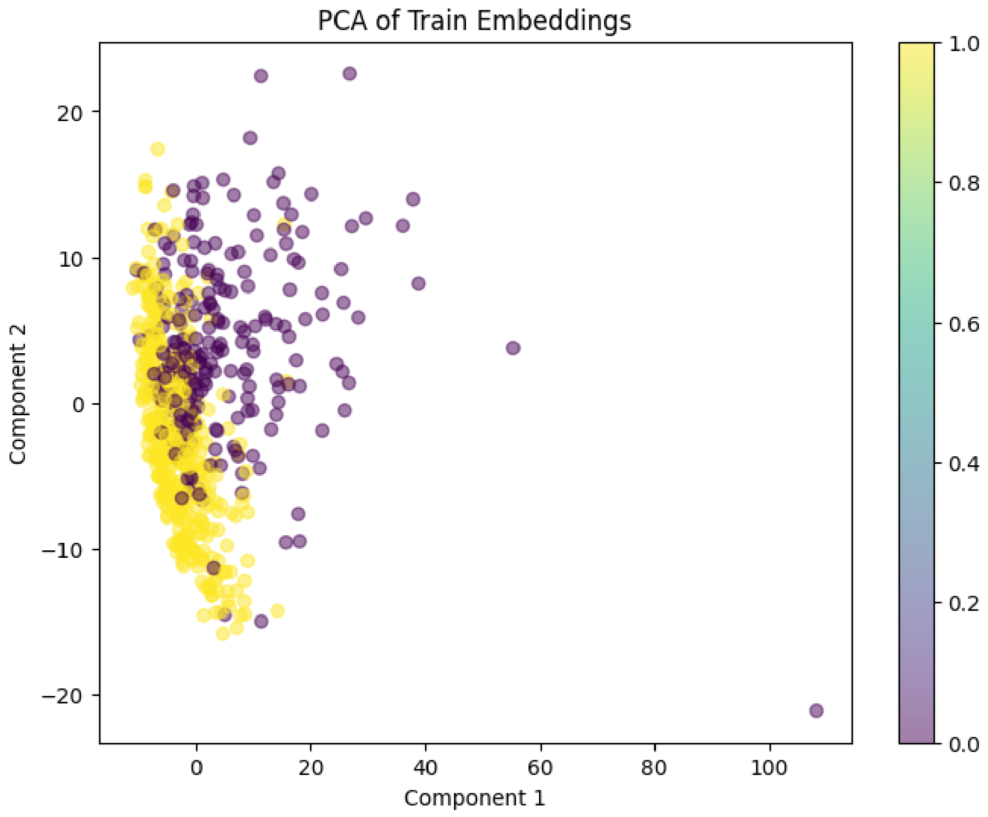


Figure 6: PCA visualization of the fine-tuning method

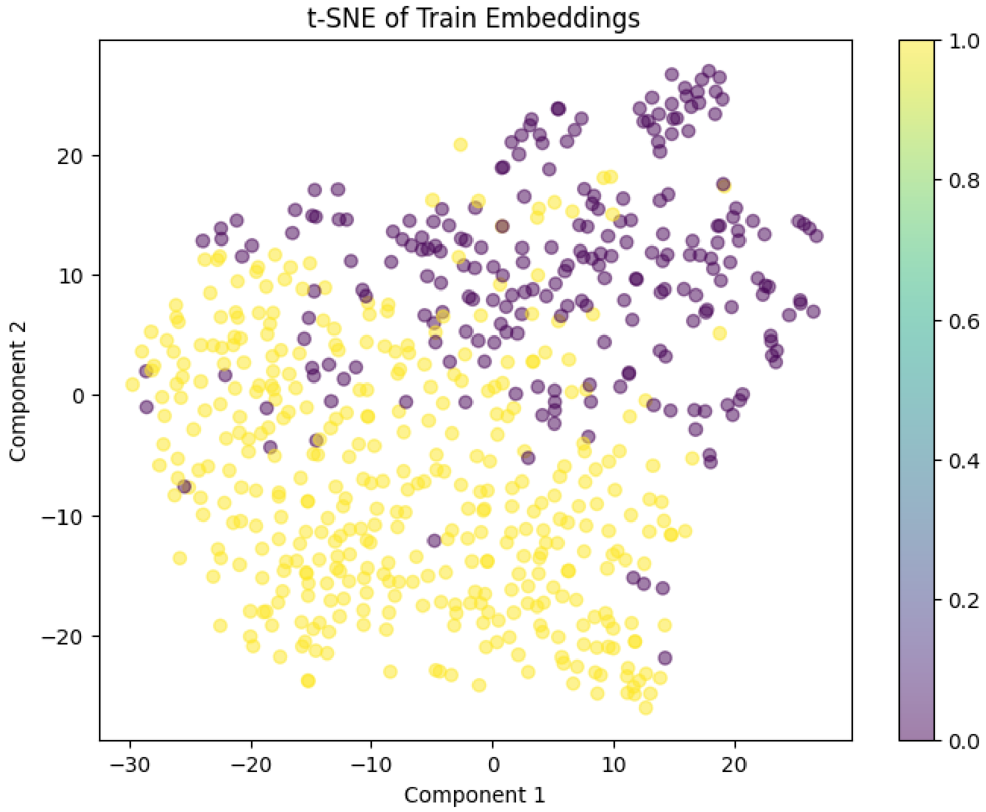


Figure 7: T-SNE visualization of the fine-tuning method