ELEC4010N Assignment 3

Pranav Gupta (20682816)

Problem 1:

Q1) In X-ray imaging, an X-ray beam, consisting only of waves with enough energy to penetrate the human body, is passed through the body. Owing to the density of the internal organs, X-rays can be absorbed or scattered - for example, X-rays can be absorbed by bones because bones contain Calcium. The remaining X-rays are recorded by a detector, forming a single image that is used for further processing. X-ray images are then processed to enhance diagnostic information, which can be achieved through dual energy decomposition - a processing technique through which a bone image and a soft tissue image can be extracted.

Q2) Filtered back projection is an analytic reconstruction algorithm for 2D images through which a sinogram (as constructed based on the detector readings during imaging) is converted into a cross-section image. It uses simultaneous equations of ray sums taken at differing angles to compute the attenuation coefficients for each angle, with which the image is reconstructed. The attenuation coefficient is simply the change in intensity when comparing the intensity of a given X-ray beam at the source and the intensity of the beam at the detector (after passing through the body). Essentially, the back-projections of different angles are added up, after which a sinogram is generated with Radon transform. The sinogram then undergoes the filtered back projection algorithm (inverse Radon transform), in which it undergoes a row-by-row fourier transformation, the output of which it is rearranged into a circle, and then which undergoes a 2D fourier transform to reconstruct the cross-section image.

However, this image is blurry because lower frequency components are sampled more densely than higher frequency components in the fourier domain, making it blurrier as it is the higher frequency components which contain the sharp edges and details. Hence, a high-pass filter is applied during the Radon transform to suppress the lower frequency components, resulting in less blurring. The addition of this filter is required in Filtered Back Projection to reduce blurriness.

Q3) Similar to X-ray imaging, Computed Tomography (CT) imaging also uses X-rays along with detectors to record the intensities of the X-rays after passing through the human body, and create a cross-section image accordingly. In this case, however, the X-ray beam can be rotated around the body to generate cross-section images of different planes (such as axial,

coronal, and sagittal plane images). With the intensity information recorded by the detectors, the signals undergo Radon transform to generate a sinogram, which is passed to the filtered back projection algorithm described above to reconstruct a “slice” - cross-sectional image. The slices are arranged to generate a 3D representation of the tissue being inspected.

Q4) CT imaging and MRI have several differences in terms of the imaging principles that are used. In CT imaging, X-rays are passed through the body at different angles to collect the respective projections, which are then used to generate the cross-section image through filtered back projection. However, in MRI, a strong magnetic field is used instead. As nuclei in the body tissue being examined are all charged (they consist of protons and neutrons), the nuclei align themselves along the magnetic field. A radio frequency pulse is then aimed towards the tissue - the pulse causes the nuclei to deviate from the alignment towards the magnetic field, and when the pulse ends, the nuclei releases the extra energy (energy from the pulse which pushed the nuclei away from the alignment toward the magnetic field) in the form of a radio-frequency wave. These RF waves are then measured as free induction decay (FID) signals, which are used to reconstruct 3D grey-scale MRImages. In short, the imaging principle is fundamentally different from CT imaging as it involves the use of a magnetic field and radio frequency pulses instead of X-ray beams. Generally, it is said that MRIs show better resolution and details compared to CT scans.

Q5) Problem Statement: Accurate 3D representations are crucial to improve effectiveness of treatments in patients - be it through better surgical planning, accurate medical diagnoses, or better patient follow-up. There are several limitations in conventional CT and MRI modalities compared to X-ray imaging, including the higher costs, higher radiation doses, limited scanning positions, and the restrictions on patients with metal devices (such as pacemakers and cochlear implants) for MRIs. Therefore, 3D reconstruction techniques of X-ray images are necessary to improve the treatment processes of different patients and to improve the accessibility of medical imaging technology to more patients.

Literature Review:

(1) X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks - (X. Ying et al., 2019)

Traditional CT imaging requires numerous X-ray projections in a full rotational body scan to generate 3D representations of the tissues being inspected. This paper proposes an alternative technique to generate the 3D representations - X2CT-GAN, a Conditional Least Squares Generative Adversarial Network (CLSGAN) which takes only two orthogonal 2D X-ray images to generate a 3D CT volume, increasing the data dimension. Trained with adversarial loss, reconstruction loss, and projection loss, the GAN is able to reconstruct complex 3D representations with relatively higher PSNR and SSIM than similar models. However, there is much doubt over the accuracy of the model, and its use is believed to be limited to niche applications such as determining the size and positioning of major organs.

(2) End-To-End Convolutional Neural Network for 3D Reconstruction of Knee Bones From Bi-Planar X-Ray Images - (Y. Kasten et al., 2020)

This paper aims to use a CNN architecture in which two bi-planar X-ray images are inputted, to return a 3D representation of the knee bone (This model was trained on knee bone images only). This model includes a segmentation network which classifies every pixel into 5 classes (background and 4 major leg bones). The model also uses a 3D Distance Weight Map that holds every pixel's distance from a bone surface, which is used in the image reconstruction. The model achieves an average dice coefficient of 0.906; however, it is limited to only knee bones.

(3) Single-image Tomography: 3D Volumes from 2D Cranial X-Rays - (P. Henzler, 2017)

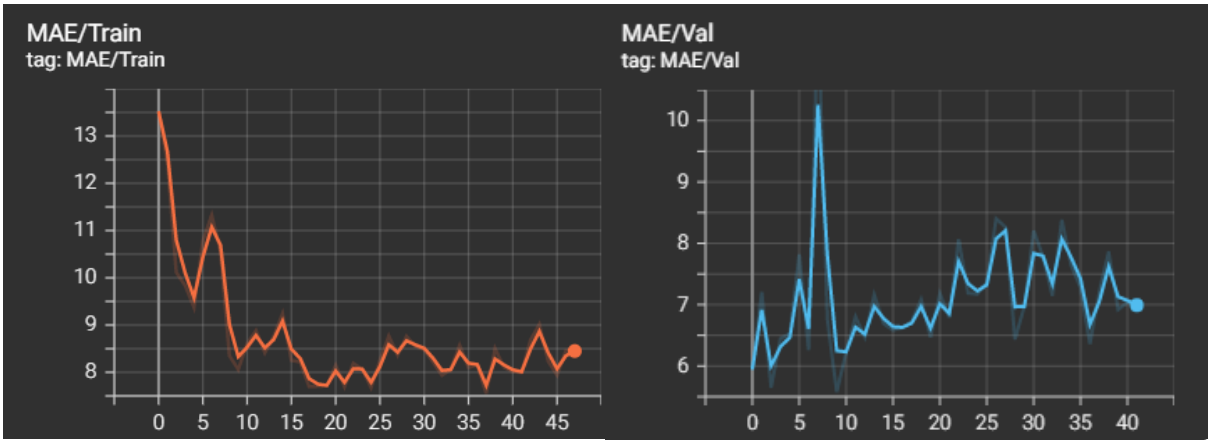
This paper presents a CNN architecture that outputs a 3D volume from a single cranial X-ray image. Unlike the papers above, this model implements residual blocks to retain spatial

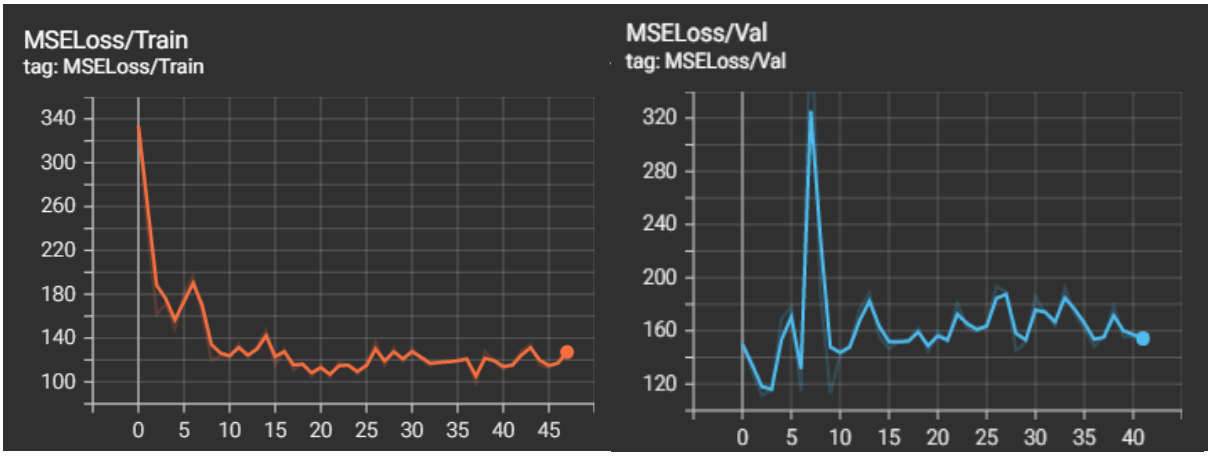
information. The model achieves better SSIM and L2 results compared to similar models. While the model was able to generate 3D representations of a skull with a single X-ray image, it may not be reliable as only one image is used. This model could be extended to include more X-ray images for better accuracy.

Problem 2:

This task aims to predict the Ejection Fraction (EF) from a dataset of ultrasound videos of beating hearts (EchoNet-Dynamic). A subset of 500 videos was used in training of the video regression model. The following presents the effects of varying loss functions and of adding supervision to intermediate layers on the overall accuracy in predicting the Ejection Fraction.

Q1) The r2plus1d_18 pretrained model from torchvision.models.video was selected for this video regression task. Videos were sampled in the shape of 64x112x112, where 64 is the number of frames chosen. The Mean-Average Error (MAE) and Mean-Squared Error (MSE) curves for both training and validation are shown below. (The MAE/Val and MSELoss/Val are from another run due to incorrect y-axes in the original Val runs)





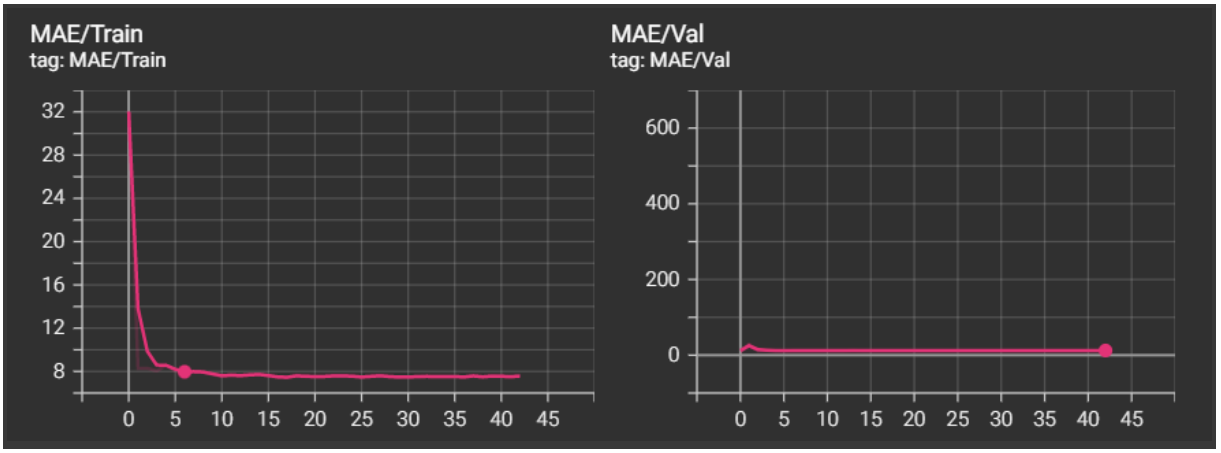
After 48 epochs, the model achieved an MAE of around 8.7 and an RMSE of around 14.3 on the test set. It is evident that the model is overfitting - more transforms and dropout layers could be included in the future. More variations of hyperparameters could be tested for optimisation, and the number of epochs could also be increased.

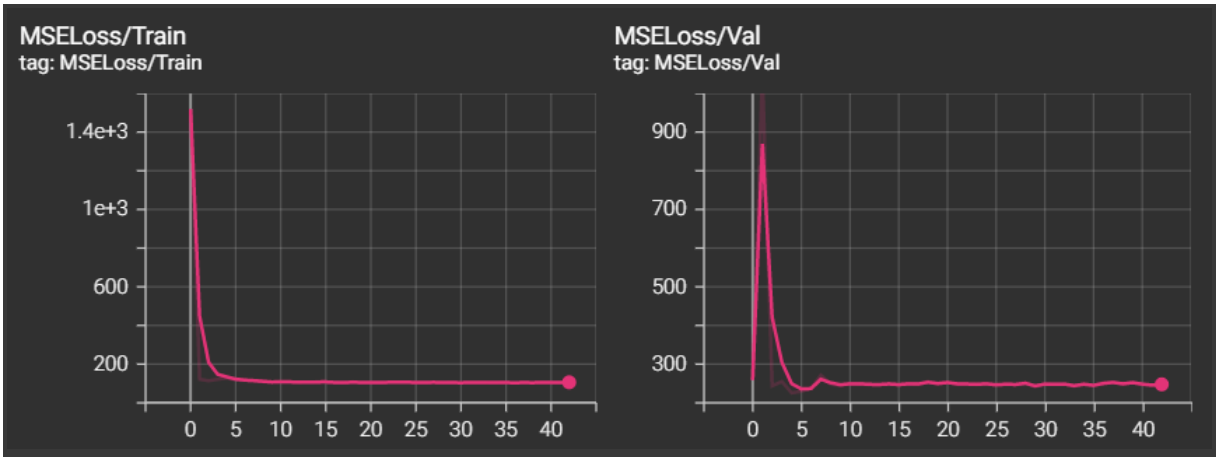
Q2) The model was trained again with different loss functions, namely the Mean-Squared Logarithmic Error (MSLE), the L1 Norm, and the Huber Loss. The performance of these loss functions was evaluated by running the model on a test set with an MSE Loss function to compute the accuracy of the trained model (Final MSE Loss). The table below presents the results.

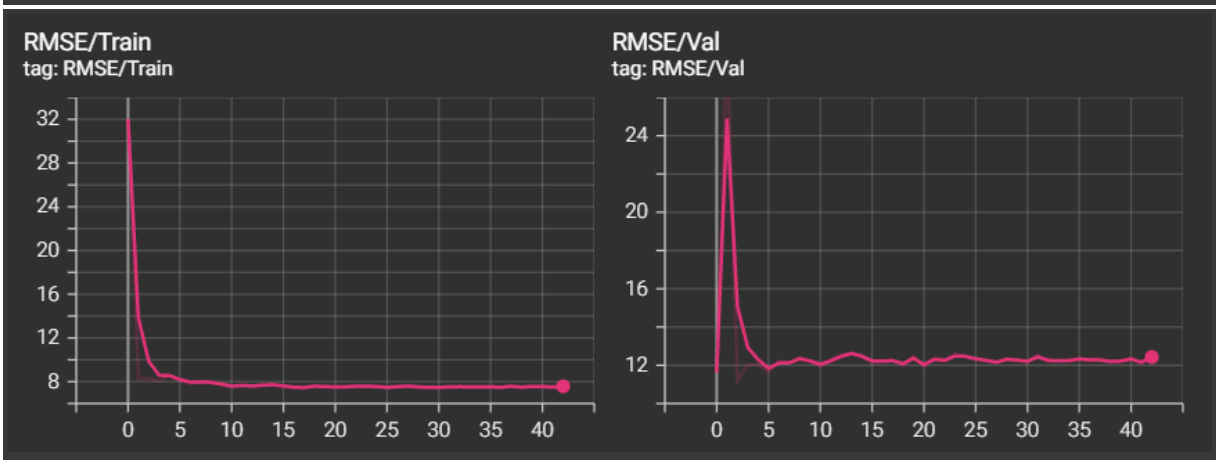
Loss Function	Final MSE Loss	Error in EF
MSE	204.64	14.31%
MSLE	158.75	12.60%
L1 Norm	238.04	15.43%
Huber Loss	356.51	18.89%

The MSLE Loss function was the best performing loss function - possibly because it implements the MSE on the log of both the predicted and targeted EF value. This function heavily penalises the weights if the predicted EF is smaller than the target EF - this function is aggressive on large errors, which is why it achieves the lowest final MSE. The Huber Loss was the worst performing loss function, possibly because it penalises large errors relatively weakly - with the default delta = 1, the function is essentially the L1 Loss - 0.5. For L1 errors within the delta, the loss function is essentially the MSE multiplied by 0.5. Hence, the Huber Loss proves to be a relatively weaker loss function. Some examples of loss curves for training and validation are attached as appendix below.

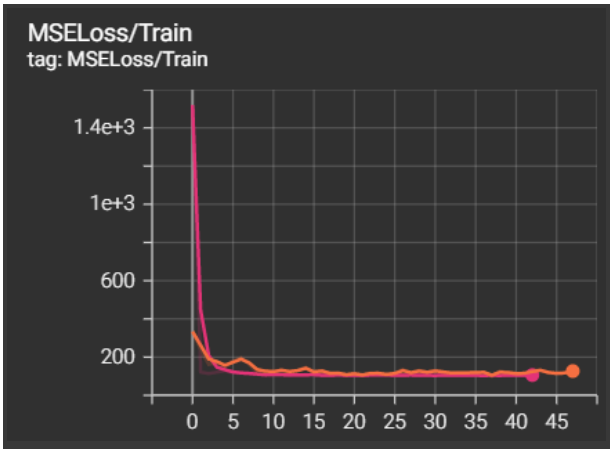
Q3) Supervision was added after the 2nd major layer of the r2plus1d_18 model (containing 4 major layers). For the supervision, a simple CNN was implemented - consisting of a 3D Conv. block, ReLU, Adaptive 3D Average Pooling block, and a Fully Connected Layer to output the predicted Ejection Fraction. The Mean-Squared Errors were computed after this layer and at the end of the r2plus18_18 model, generating two losses, both of which were used for back-propagation. The plots below show the Mean-Average Errors (L1 Loss), Mean-Squared Errors, and the Root-Mean-Squared Errors, for both training and validation.


MAE/Train
tag: MAE/Train
32
28
24
20
16
12
8
0 5 10 15 20 25 30 35 40 45
MAE/Val
tag: MAE/Val
600
400
200
0
0 5 10 15 20 25 30 35 40 45


MSELoss/Train
tag: MSELoss/Train
1.4e+3
1e+3
600
200
0 5 10 15 20 25 30 35 40
MSELoss/Val
tag: MSELoss/Val
900
700
500
300
0 5 10 15 20 25 30 35 40


RMSE/Train
tag: RMSE/Train
32
28
24
20
16
12
8
0 5 10 15 20 25 30 35 40
RMSE/Val
tag: RMSE/Val
24
20
16
12
0 5 10 15 20 25 30 35 40

For reference, the plot below compares the Mean-Squared Errors over epochs during training between the original model (orange) and the model trained with supervision (pink). The curves may seem to converge at the end; however, that is only due to the y-axis being large to cater for the initial large loss values - the model with supervision maintained a lower MSE loss throughout (excluding the initial stage). Adding supervision definitely helped the model achieve a higher accuracy in predicting the Ejection Fraction, and also helped the model converge faster (despite not being pre-trained as the original model.)



Appendix:

Link to pdf of papers reviewed (in order):

https://arxiv.org/pdf/1905.06902.pdf

https://arxiv.org/pdf/2004.00871.pdf

https://arxiv.org/pdf/1710.04867.pdf

Examples of Training and Validation Loss Curves of different Loss Functions