Automated measurement of skin, fat and muscle thickness from ultrasound images

Sebastian Janampa
sebastian.janampa@utec.edu.pe

1 Methodology

1.1 Metrics

Metrics are used to compare different models and to determine which the best model is. In this project two metrics are used the*Inter-Observer Error* and the *Intra-Observer Error*. In both cases, the *Mean-Absolute Error* (MAE) as the metric error. To test both metrics of measuring the thickness of the skin, muscle and fat, 5 subjects were randomly selected, each of them has a dataset of 48 anatomical trials [1].

1.1.1 Inter-observer error

The *Inter-Observer Error* describe the variances between measurement done by two different observers. In the project, this error is the maximum error that the network has to achieve.

1.1.2 Intra-observer error

The *Intra-Observer Error* is similar to the *Intra-Observer Error* but it describes between the same observer. This error is the minimal known error that the network could achieve because there is not previous literature about neural networks to measure the thickness of the project's interest tissues.

1.2 First phase

1.2.1 Data normalization

The first network, model A, is built to determine the efficiency of normalizing the output data. Model A consists of five convolutional layers with a kernel size of (3,3), padding and a He normal initializer. The layers' activation function is the Rectifier Linear Unit(ReLU). Max-Pooling layers, whose Kernel size is (2,2) and stride equals to 2, is applied after every convolutional layer. Its loss function is the *Mean-Square Error* (MSE). A previous visualization of the network architecture is in Fig. 1.

Technique	Formula
z-score	$x' = (x - \mu)/\sigma$
linear scaling	$x' = (x - x_{min})/(x_{max} - x_{min})$
decimal scaling	$x' = x/10^j$

Table 1: Normalization techniques. μ and σ are the mean and the standard deviation of x, respectively. j is the minimum integer that satisfies the condition $|x'| \leq 1$

Because of there are multiple output variables (skin, fat and muscle thickness) which own different scales, it is difficult for the weights to tune themselves to every scale. A way to afford this problem is to normalize the data to have a better weight initialization . The normalization techniques used are: z-score, linear scaling and decimal scaling. The formulas of each technique are shown in Table 1.

1.2.2 Costum loss function

Because of the data is normalized, the network will not consider that an error of one tissue is more significant than other. To solve this, a weighted matrix is used. This matrix contains the median value of every tissue.

The second network, model B, has the same architecture than model A (Fig. 1) but it has different loss function based on the *Mean-Square error*. Its

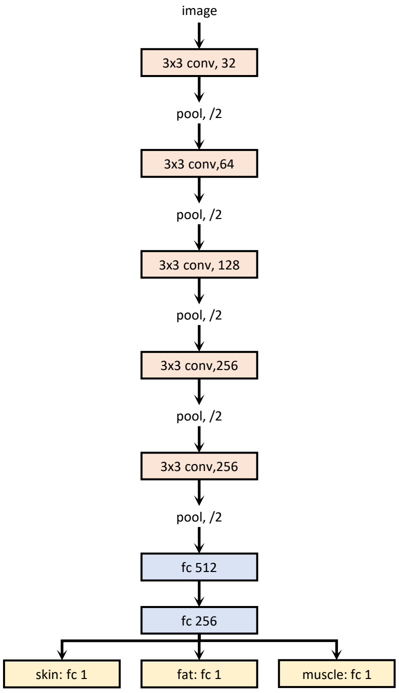

Figure 1: Architecture of model A for experimenting with normalizatión and costum function

formula is:

$$L(y, \hat{y}) = \frac{1}{3m}(W \cdot \sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2) \qquad (1)$$

where y and $\hat{y}$ are the true values and predicted values whose size is $(m$,3) where m is the number of samples. W is the weight matrix with a size of (1,3). The number 3 in the denominator is for the number of tissues.

1.2.3 Categorical data

To validate the relationship between categorical data and the thickness of the tissues and to determine if whether or not adding the categorical data increase the performance of the network, a third network, model C, is built. This model has the same number of convolution layers with the same parameters as model A and B. More fully connected layers are added to introduce the categorical input data. An overview of the structure is shown in Fig 2.

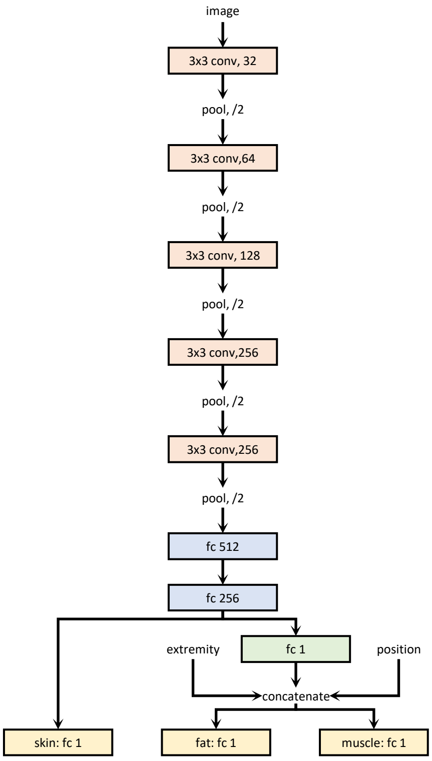

Figure 2: Architecture of model C for experimenting with categorical data. This model is similar to model A. The differences are the green block and the concatenate layer that let to add the categorical input data.

1.3 Second phase

Improving the error of the networks from Sec. 1.2 is the objective of this phase.

1.3.1 Deeper models

Khan, Sohail, Zahoora, *et al.* show in [2] the advantages, the main contributions, the top-5 error rates, etc. of different networks architectures. The networks chosen for this project are Inception-V3 [3], ResNet [4], [5], WideResNet [6] and DenseNet [7]. Different models are created based on these networks to determine which are the two models with the lowest error. The depths of these models are not greater than 30 because of the computational cost and the time required to train them. The models architectures are depicted in Fig. 3.

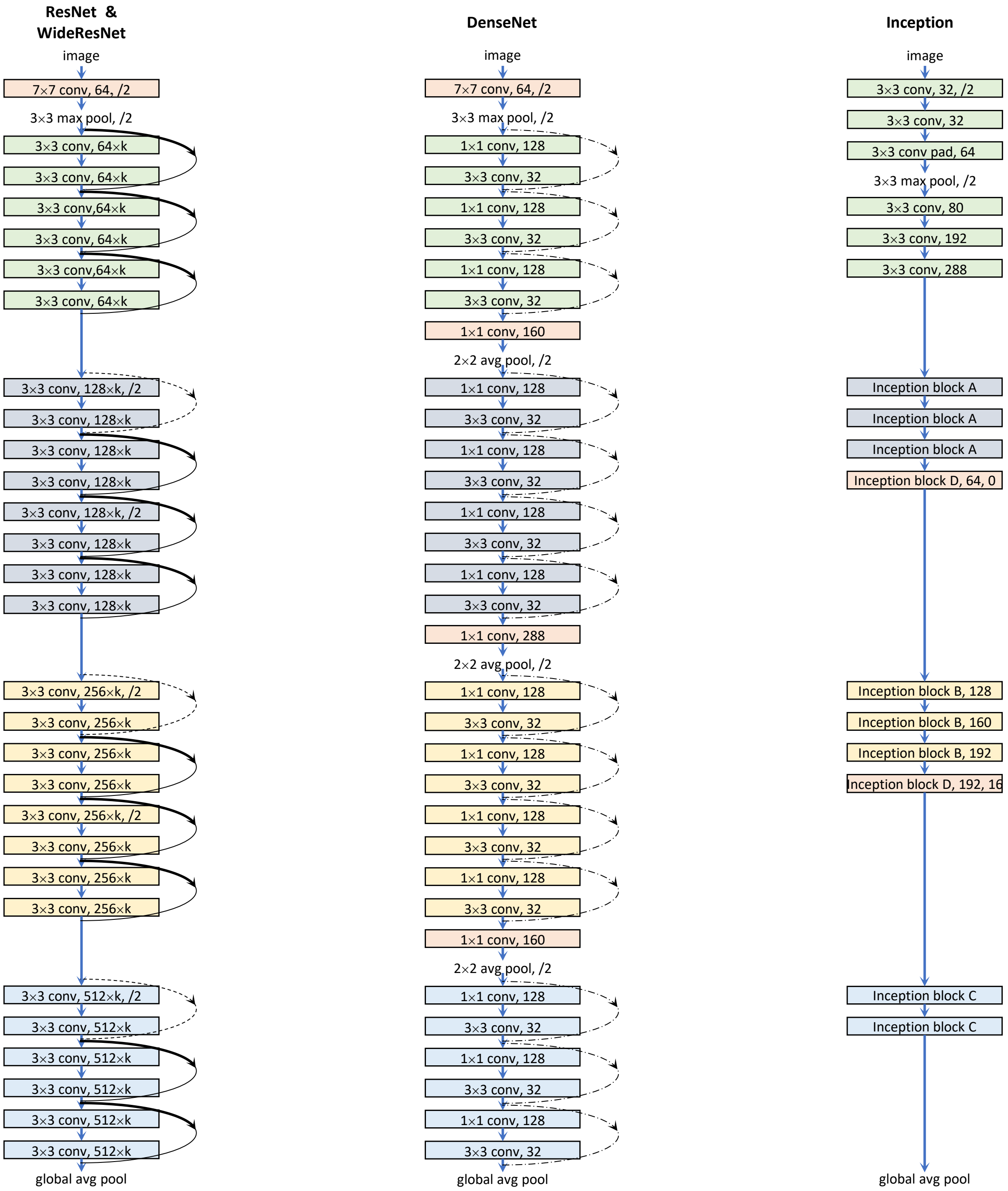

Figure 3: Deeper model architectures. In the ResNet & WideResNet model, k is the width factor which increases the number of channels. The value of k is 1 and for ResNet and WideResNet, respectively. Solid lines represent the addition operation and the dotted shortcuts are projections with a stride of 2 to increase to match the number of channels. In DenseNet, the lines represent the concatenation operation. The blocks of the first models follow the structure BN-RELU-CONV and include padding. For inception architecture,the blocks structure is CONV-BN-RELU and do not include padding except it is indicated. The Inception blocks B and D have one and two parameters, respectively, which are explained in Fig. 4

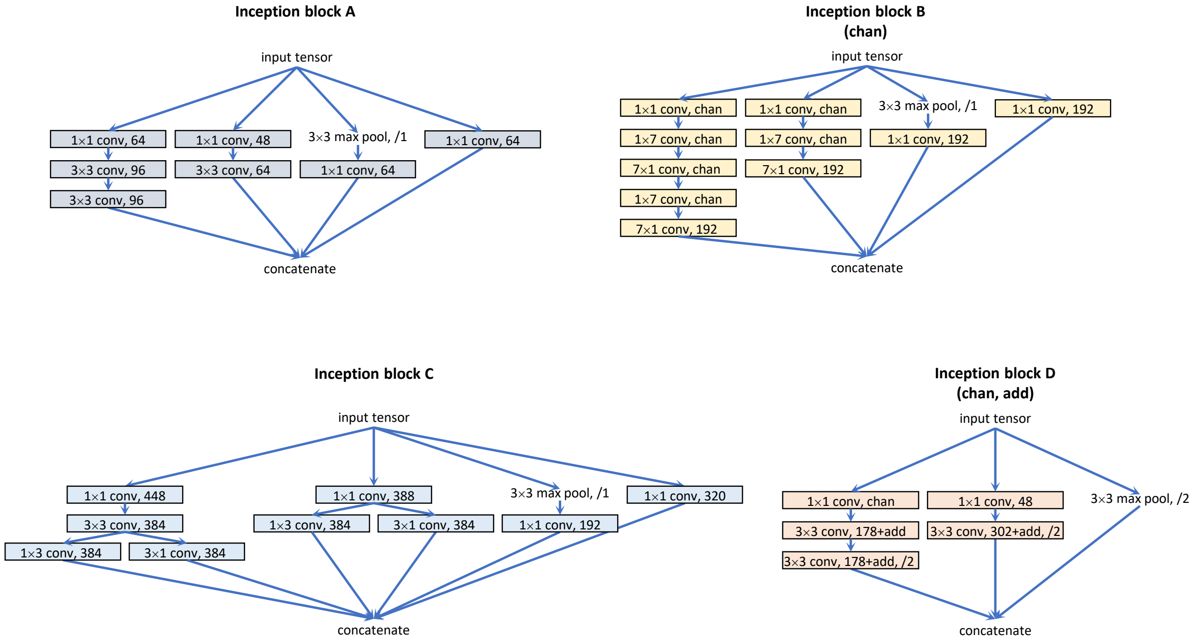

Figure 4: The Inception blocks are proposed by Szegedy, Vanhoucke, Ioffe, *et al.* in [3]

1.3.2 Squeeze & Excitation Block

Hu, Shen, Albanie, *et al.* introduce Squeeze & Excitation(SE) block and its variants in [8].This block performs channel-wise feature recalibration by explicitly modelling interdependencies between channels to selectively emphasise important features and suppress less useful ones. SE block consist of two parts: the squeeze and excitation. First, it *squeeze* global space information ($H \times W \times C$) into a channel feature ($1 \times 1 \times C$). Finally, it uses a simple gating mechanism with a sigmoid function to learn a non-mutually exclusive relationship between channels and to map the input from *squeeze* to a set of channel weights. A graphical representation is shown in Fig. 5. SE-PRE block with a ratio of 16 is included in the two best models from Sec. 1.3.1.

1.3.3 Best Model 1

1.3.4 Best Model 2

1.3.5 Medical networks

The models shown above are used to classify no-medical images. For this reason, U-Net [9] and U-net^{++} [10] are used. Moreover, another difference is that these networks are used for segmentation.

1.3.6 Fully-connected layers

Basha, Dubey, Pulabaigari, *et al.* describe the impact of fully-connected(FC) layers in the convolutional neural networks [10]. They propose a method which consists in, initially, train a model with a single FC layer (output layer). Then, adding a new FC layers before the output layers whose number of neurons is the number of output neurons or all powers of 2 up to 4096. For instance, for the project, it could add a new FC layer with 3, 4, 8, 16, ..., 4096 number of neurons. The next step is to add one more FC layer before the recently added FC layer. Its number of neurons could be the number from neurons of the previous FC layer up to 4096.

1.3.7 Augmented training data

A key factor of this project is the relatively small set of 4799 images. Data augmentation is a technique to increase the size of the training data. By rotating the images slightly, a new image can be generated. In this project, horizontal flipping and translation,

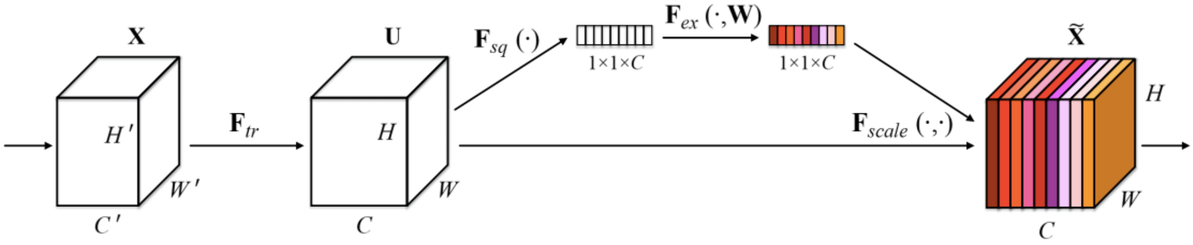

Figure 5: Squeeze & Excitation block. The image is from [8]

and rotation are applied to the image.

$$maxRot = 30^\circ$$
$$flipping = horizontal$$
$$maxTrans = 20pxs$$

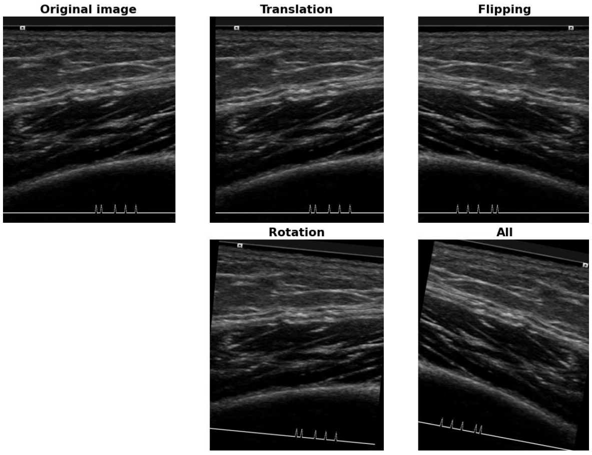

Figure 6: Examples of augmented training data. A random flipping, translation and rotation are applied to the images to increase the size of the training dataset.

2 Results

In this section, the results of the experiments in Section 1 are presented. The measure of accuracy is the *Mean-Absolute Error*. Its formula is

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |y_j^{(i)} - \hat{y}_j^{(i)}|$$

where m is the number of samples, $y_j^{(i)}$ is the real value of the sample i of tissue j and similarly $\hat{y}_j^{(i)}$ is the predicted value by the network.

2.1 Metrics

2.1.1 Inter-observer variability

The inter-observer error is calculated as MAE using the difference between two measurements of different experts on the 240 anatomical trials (48 anatomical trials per patient). The results are shown in Table 2.

These measurements serve as benchmarks for evaluating the networks and comparing them with the manual annotation process.

2.1.2 Intra-observer variability

Similar to the inter-observer error, the difference between two measurements of of the same expert on the 240 anatomical trial was used to get the intra-observer error. The results are shown in Table 2.

These measurements serves as the lowest error expected from the networks. These are not used for evaluating metrics.

Error	MAE Skin	MAE Fat	MAE Muscle
Inter-observer[1]	0.36 ± 0.38	0.78 ± 1.10	0.65 ± 1.11
Intra-observer[1]	0.16 ± 0.17	0.49 ± 1.19	0.48 ± 1.19

Table 2: MAE for Inter- and Intra-observer error of every tissue.

2.2 First phase

This subsection reports the results of the networks test for phase 1 of normalizing the data and adding categorical inputs to the network.

2.2.1 Data normalization

The results of normalizing the output data are shown in Table 3. Every results (including the data without normalization) have a great performance on measuring the skin thickness. The errors are less than the the inter-observer error of the skin (see Table 2). For the remaining tissues, the error are much greater than the inter-observer error, which is expected because of the architecture of model A was created to validate if applying normalization improve the performance, not to get an acceptable performance. All the networks show greater values in the validation set than in the test set for the fat and muscle which could indicates that measuring the thickness of the fat and muscle are more difficult in the validation set.

The lowest test errors are from the data without normalization and with z-score normalization. The model with z-score normalization drastically overfits the data for the tree tissues. The model without normalization has the same issue but only for the fat and muscle. On the other hand, linear scaling and decimal scaling have the greatest errors. For fat, they slightly overfit the data and their errors are closed to each other with a difference of 0.2 units approximately. For the muscle, they also slightly overfit the data in the test set, and the error of the decimal scaling in this set is greater for almost 1.2 units. In the validation set, the decimal scaling technique is better than linear scaling by a difference of approximately 2 units. For these reasons, z-score and decimal scaling are considered for the next step of the project.

References

[1] E. E. Neumann, T. M. Owings, T. Schimmoeller, T. F. Nagle, R. W. Colbrunn, B. Landis, J. E. Jelovsek, M. Wong, J. P. Ku, and A. Erdemir, "Reference data on thickness and mechanics of tissue layers and anthropometry of musculoskeletal extremities," *Scientific Data*, vol. 5, no. 1, Sep. 2018. DOI: 10.1038/sdata.2018.193.

[2] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020, ISSN: 1573-7462. DOI: 10.1007/s10462-020-09825-6. [Online]. Available: https://doi.org/10.1007/s10462-020-09825-6.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Dec. 2015. arXiv: 1512.00567 [cs.CV].

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Dec. 2015. arXiv: 1512.03385 [cs.CV].

[5] ——, "Identity mappings in deep residual networks," Mar. 2016. arXiv: 1603.05027 [cs.CV].

[6] S. Zagoruyko and N. Komodakis, "Wide residual networks," May 2016. arXiv: 1605.07146 [cs.CV].

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," Aug. 2016. arXiv: 1608.06993 [cs.CV].

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," Sep. 2017. arXiv: 1709.01507 [cs.CV].

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," Jul. 2018. arXiv: 1807.10165 [cs.CV].

[10] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," Jan. 2019.

	MAE Skin			MAE Fat			MAE Muscle		
Model	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Without normalized	0.272 ± 0.17	0.267 ± 0.16	0.319 ± 0.18	0.706 ± 0.44	1.903 ± 1.39	1.893 ± 1.34	0.609 ± 0.37	2.413 ± 1.62	2.231 ± 1.43
Z-score	0.076 ± 0.04	0.242 ± 0.16	0.254 ± 0.15	0.784 ± 0.49	1.942 ± 1.36	1.876 ± 1.33	1.678 ± 0.98	2.861 ± 1.84	2.884 ± 1.79
Linear scaling	0.226 ± 0.14	0.271 ± 0.17	0.254 ± 0.17	2.091 ± 1.39	2.646 ± 1.83	2.443 ± 1.72	4.279 ± 2.28	5.308 ± 2.93	4.466 ± 2.56
Decimal scaling	0.310 ± 0.19	0.312 ± 0.19	0.338 ± 0.20	2.431 ± 1.65	2.713 ± 1.80	2.670 ± 1.81	2.723 ± 1.77	3.453 ± 2.20	3.094 ± 2.09

Table 3: MAE for experiment with normalized input on Model A.

DOI: 10.1016/j.neucom.2019.10.008.
arXiv: 1902.02771 [cs.CV].