TECHNISCHE UNIVERSITÄT DRESDEN

INSTITUTE OF SOFTWARE AND MULTIMEDIA TECHNOLOGY
CHAIR OF COMPUTER GRAPHICS AND VISUALIZATION LAB
PROF. DR. STEFAN GUMHOLD.

Multi-modal Medical Image Fusion using Convolutional Neural Networks

Bachelor thesis

Submitted to the faculty of Computer Science at the Technische Universität Dresden in partial fulfillment of the requirements for the degree of
Bachelor of Science

Cai Badal Regàs
(Born 24. May 1996 in Barcelona)

Supervisor: Prof. Dr. Stefan Gumhold
Tutor: MSc. Nishant Kumar

Dresden, December 17th, 2018

Task assignment

- An extensive literature research on existing methods related to image fusion with emphasis on medical field.
- Implement a convolutional neural network for medical image fusion on the given two modalities:
  1. Implement a convolutional neural network architecture with four or more hidden layers and decide upon the size of kernel filters along with the choice of activation function, optimization function, hyperparameters and so on.
  2. Implement structured similarity index (SSIM) as the loss function on the network‘s output to maximize the similarities between fused information and the input images.
  3. Perform training of the network architecture and optimize the network.
  4. Perform testing of the network architecure with a pair of input images and extract the resultant fused image.
  5. Implement the task in python using frameworks such as numpy, tensorflow etc.
- Detailed evaluation of the developed solution:
  1. Qualitative evaluation by comparing the fused image with the outputs of other image fusion approaches such as Guided Filtering, Weighted averaging.
  2. Quantitative evaluation of results in terms of accuracy: Implement of evaluation metrics such as Feature mutual information (FMI), SSIM etc on the developed solution and compare the results with Guided Filtering, Weighted Averaging etc.

Statement of Authorship

Herewith I declare that I am the sole author of the thesis named:

Multi-modal Medical Image Fusion using Convolutional Neural Networks

which has been submitted to the study commission today. I have fully referenced the ideas and work of others, whether published o-published. Literal or analogous citations are clearly marked as such.

Dresden, 17th December 2018

Cai Badal Regàs

Abstract

Multimodal image fusion merge images of different modalities into a single image which contains greater information than any of the input images. In medical field, multimodal image fusion plays a crucial role in providing medical practitioners sufficient information about the input images for clinical purposes. In recent years, deep learning (DL) based image fusion has achieved remarkable breakthroughs and state of the art results owing to strong capability in feature extraction. One of the challenge that is common in the DL based image fusion is the unavailability of ground truth. Therefore, this thesis aims to develop a convolutional neural network based medical image fusion approach on two different 2D modalities without any groundtruth information. The end to end learning framework proposed in this work presents a completely new approach to deal with the problem of multimodal medical image fusion where we obtain a fused image as an output of our network. Our results shows that we achieve promising results after comparing the method with other medical image fusion approaches available in literature.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Image fusion

Multimodal image fusion is the process of integrating different images with complementary information into a fewer number, usually a single one, so that the fused image is more suitable for either human visual perception or computer-processing tasks, such as segmentation, feature extraction and object recognition. The main goal is to obtain a single fused image that can provide better and more relevant information than any of the source images on their own.

Image processing applications increasingly demand the development of image fusion due to the constraints that a single image sensor has, such as optical limitation for a single focus, improper image capturing or lack of clarity and quality. [37] By implying the integration of multiple images acquired by multiple sensors that can provide little segments with clarity and quality, we can get a better perspective of a scene that contains more information. The fused image should contain the complementary as well as the common features of the source images, giving superior information for both subjective and objective analysis.

A clear example of the utility of image fusion can be seen in Figure 1.1, that illustrates the combination of two brain medical imaging modalities such as CT and SPECT, that will be more deeply explained in another section. The images are obtained from Harvard Medical School's Attlas dataset[1] of a 42 years old woman. By the fusion of those we obtain a single image that combines the information of both modalities and give as a more accurate information of the brain as shown in the third image of Figure 1.1. [30]

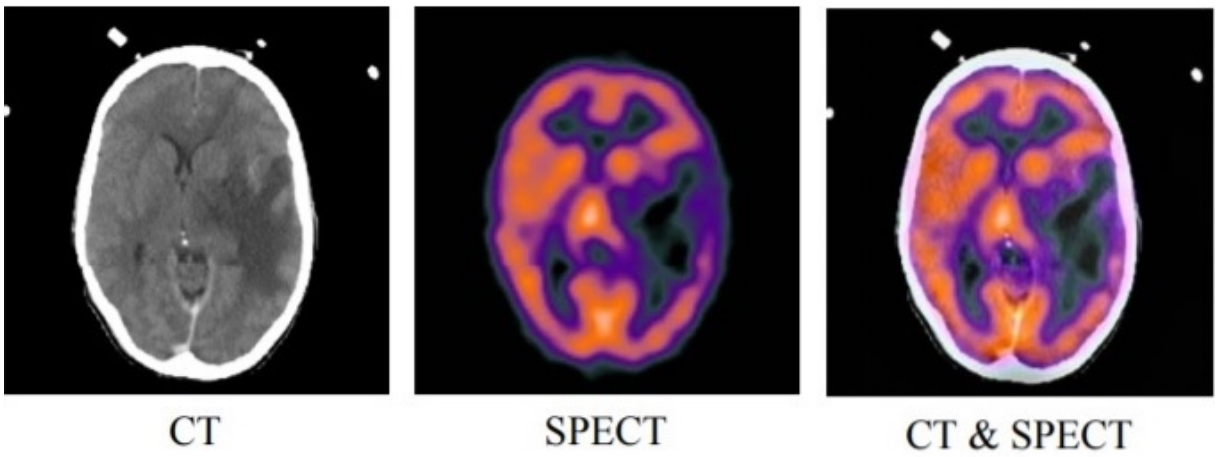

Figure 1.1: Example of CT and SPECT brain image fusion.

One of the main obstacles that image fusion finds is to get an objective evaluation metric of the output images, since there is no ground truth available. Furthermore, not always the source image are available for the evaluation of the fusion and therefore the metrics are divided in two main groups depending on whether the source image are accessible or not.

The most popular metrics belonging to the first group are only used for image fusion problems and can be grouped into four categories: the information theory based metrics, the image feature based metrics, the image structural similarity based metrics, and the human perception based metrics[33]. The metrics for the group where the source images can not be used to obtain fusion quality metrics are only based on standard image quality metrics, like standard deviation, spatial frequency or entropy.

1.2 Types of image fusion

Depending on the goal and the kind of images that are desired for the applications, image fusion can be briefly divided into five sub-categories: multiview fusion, multimodal fusion, multitemporal fusion, multifocus fusion and fusion for image restoration. [23]

a) **Multiview fusion:** the source images are from the same modality and taken at the same time but from different places or even different conditions or environments. These different images are combined to get complementary information from distinct views.

b) **Multimodal fusion:** a very popular fusion for medical purposes, where different modalities of images are used to decrease the amount of data while emphasizing band-specific information. This thesis is going to be based in medical applications of this type of fusion, which doesn't try to decrease the amount of data but instead tries to take as much information as possible from the input images into the fused images.

c) **Multitemporal fusion:** the images, usually from the same modality, are taken in the same conditions but with a temporal delay between them. The main goal is to detect the changes by subtraction of the images.

d) **Multifocus fusion:** the images are taken with different focuses and they are fused so that the final image has all regions focused. A very popular approach is the wavelet transform to identify the regions in focus and combine them together. This approach can be seen as a typical binary classification problem. By detecting the focused and unfocused regions, it is possible to approximate as well a 3D map of the precise position of the scene.

e) **Fusion for image restoration:** each of the source images, from the same modality and scene, consist of a true and degradation part. The combination of different blurred images can lead to deblurred and denoised version.

According to [39], we can divide image fusion algorithms in two groups depending on the type of fusion performed between spatial or transformed domain. The first method directly deals with the the pixels of the source images to get the result. The spatial domain fusions can be as simple as averaging or minimum/maximum selection. A more successful method has been Principal Component Analysis (PCA), which is a technique used to reduce the dimensionality of data to transform it into more suitable for analysis, or Intensity, Hue and Saturation (IHS).

In the transform domain image fusion, the image is represented in the frequency domain and represented with coefficients that represent features. A decision map is required to select the desired coefficients of the transformed image and by applying the inverse transform to the mapping we obtain the fused image. One of the most popular algorithms is the Discrete Wavelet Transform [7], where the image is divided in different frequency bands.

1.3 Medical applications for image fusion

Image fusion is widely applied to medical issues for numerous purposes addressing body, organs or even cells images. Since imaging techniques allow a quantitative evaluation of the images under judgment, it helps to take objective decision on diagnosis and analysis. A very popular approach of image fusion is to get high resolution images. However, the growing method of multimodal image fusion, has opened a new window of infinite possibilities for medical analysis applications[15]. This application gives doctors more truthful information and even reveal features or abnormalities that could not have been seen by the human being, which can often lead to achieving a more precise diagnosis and treatment.

For medical purposes there are plenty of used imaging modalities, depending on the desired type of information. Computed tomography (CT) has the ability of providing a visualization with less distortion of bones, implants and other dense structures. Magnetic resonances (MR) are used to get a better view of soft tissues while positive electron tomography (PET) can show information about how tissues and organs are functioning. Single-photon emission computed tomography (SPECT) provides better information on blood flow activity with low spatial resolution. Ultrasound images provides of high frequency waves to build images of mainly non bony organs.

An important focus of medical image fusion applications has been the brain. The briefly exposed imaging methods expose several features that are crucial for diagnoses that are otherwise not captured by human-like sensory mechanisms. A numerous list of applications of image fusion applied to brains can be seen in [22]. Many papers have been published involving image fusion work related to the brain with very different applications: image guided neuro-surgery [12], classification of abnormal tissues[14], segmentation of brain tissues [28], 2D–3D registration of brain images [18] or quantification of brain tissue volumes [9] among many others.[22]

The elevated rate of breast cancer in women has given breast image fusion study a big relevance. A fusion that has shown growing accuracy in diagnosis is the combination of the information provided by X-ray computed tomography and PET. Many papers (e.g. [43] or [40]) have shown success techniques for this diagnosis. Other organs like prostate [8], lungs [38] or liver have proved to get remarkable improvements with image fusion techniques.

1.4 Future of Image Fusion

Medical applications for image fusion has shown a remarkable growing in the last years. In Figure 1.2 we can observe the growing tend in published papers on issues involving image fusion between the years 2000 to 2017. We can clearly see an increasing tendency on the publication of image fusion papers, with a great amount involving medical applications.

This increase clearly shows that image fusion is an active research area to be exploited, with many variants and opened fronts that deserve to be studied. Since a common application of image fusion is surveillance, there is a need to find more efficient algorithms for the fusion in order to be able to perform image fusion in real-time. For examlpe, any outdoor uses of image fusion are still not as robust as it could be, with changing conditions, noise or exposure problems.

Noise in image fusion is one of the fields that might get important interest in future research, since there are lots of noise types that affect fusions. Denoising procedures have not taken an important role in image fusion and might be a subject of interest. [16]

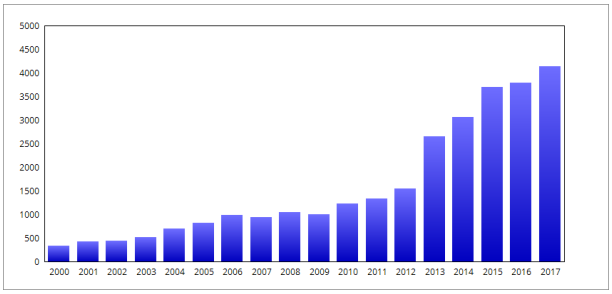

Figure 1.2: The number of publications in image fusion obtained from the Web of Science [5]

Regarding medical image fusion, a really precise registration of the images is required, specially for the fact that mainly the images are from different modalities. Additionally, objective evaluation still is a challenging issue, mainly because the evaluation of the success of a fused image is extremely different to any other kind of evaluation. Evaluating the efficiency and the help that a fused image provides is way harder than the assessment of the quality of an image.

2 Related work

Deep learning has shown an amazing growth in the last years due to the high effectiveness of its algorithms. In this section, some other DL algorithms that have been applied to image fusion will be presented and some related work on the focus of the thesis (i.e. convolutional neural networks) will be shown.

Apart from the already mentioned lack of objective evaluation metric for image fusion problems, there are other aspects where image fusions find difficulties and are worthy to study, for example: the lack of effective image representation, the desire of finding the best suitable transform and fusion strategy and the limited existence of mapping that characterize the relationship between source and targeting images. In many computer vision problems, deep learning (DL) has shown state-of-art results for its capability in feature extraction and data representation. Some of the popular DL approaches to image fusion showing high effectiveness results in different image fusion problems have shown to be convolutional sparse representation, stacked autoencoders and convolutional neural networks. [33]

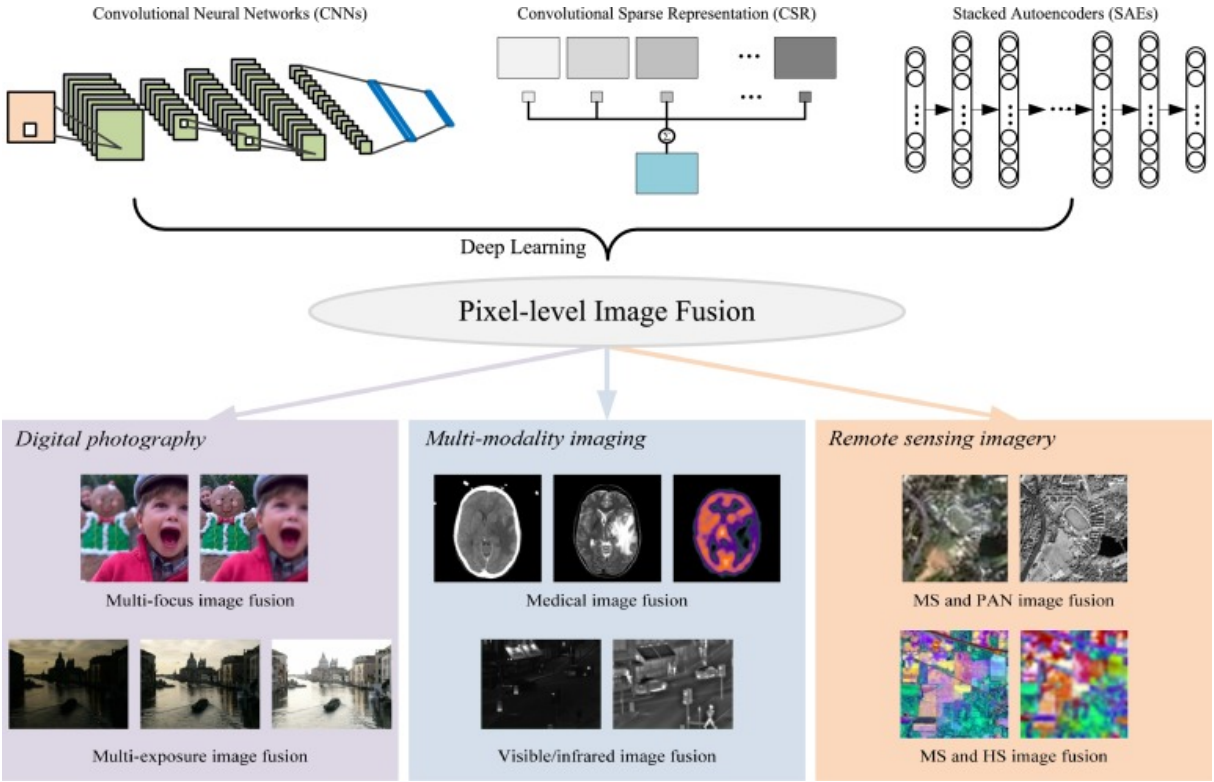

Figure 2.1: Some Deep Learning algorithms and applications from [33]

2.1 Convolutional Sparse Representation (CSR)

This concept was born from the deconvolutional networks proposed by Zeiler et al [46]. It is mainly based on sparse representation (SR), which is a popular signal modeling technique that has achieved

great success in image fusion. It is a transform approach that has provided better results than the conventional multiscale transforms technique. [31]

This method was first applied by Yang and Li, where a sliding window technique divides the images in overlapping segments that are decomposed using orthogonal matching pursuit (OMP) algorithm. After the reconstruction, the final image is obtained by combining the reconstructed patches into a single image.

Therefore, this multi-valued patches are not optimal compared to the full image. In order to improve the approach, convolutional sparse representation was developed, which is computed in a model in order to get a single-valued and optimized representation of the full image. Furthermore, CSR technique has a crucial property in many image fusion applications representation approach, which is shift-invariance representation.

2.2 Stacked autoencoders (SAEs)

An autoencoder is a layer of neural network that by applying backpropagation, tries to learn an approximation to the identity function to replicate its input at its output. The network parameters are optimized using an appropriate cost function. The main aim is to find the correlation among high-dimensional data to develop better feature representation of it. The SAEs simply consist of the combination of multiple layers of autoencoders which the outputs of each layer is wired to the inputs of the successive layer. [33]

2.3 Convolutional Neural Networks (CNN)

CNNs are architectures similar to basic neural networks with the difference that the layers are not fully connected, or at least not all of them. In this case, the neurons are connected to adjacent ones with convolution operations, which drastically reduces the number of parameters to be learned. The convolutional layer is then followed by a non linear layer, for example a pooling or max-pooling operation, which also reduces the computational cost and the number of features.

In general terms, CNNs can learn very effectively the desired features from the training data without manual help. The facility to correctly mediate between different types of signals makes CNNs a great tool for multimodal image fusion and since the architectures are very variable and adaptable, it can be a good solution to many variety of classification and regression tasks. In [30], a fusion scheme is presented in four steps. Firstly, the source images are feed into the convolution neural network that returns a weight map. Then the images are decomposed to Laplace pyramids with l levels. Finally, the coefficients of the l levels are fused and the laplacian pyramid is reconstructed and the fused image is restored.

CNNs are really flexible and can involve any kind or architectures desired. This kind of networks have shown state-of-art in plenty of problems, and it is interesting to see that the disciplines can be really different. As an example, convolutional neural networks has been used for large scale visual recognition or image understanding by academia and industrial behemoths such as Google, Facebook and Baidu [25].

The success of CNNs can be clearly seen in [41], where the authors present a framework that was awarded of the localization task of the ImageNet Large Scale Visual Recognition Challenge 2013. The network is able to detect, locate and classify objects with state-of-arts results. The approach consists multiscale and an sliding window with the novelty of predicting the object boundaries that enhances the detection accuracy. A common problem in computer vision is finding a similarity function for the comparison of

image patches. In [45] an approach to this problem with CNNs is presented. The network learns the similarity function directly from the raw pixels. Diverse and different architectures are tested and show more consistent results than the state-of-art, with special attention on that the 2-channel-based ones obtained better results than any others.

Another application for image fusion using Convolutional neural networks, as already mentioned before, is multifocus image fusion, where different images of the same scene with different focuses are integrated to obtain a better quality fused image. [32] presents a method based on a convolutional neural network with the main novelty that the activity level measurement and fusion rule are jointly created. In the same paper, both visual and objective results are demonstrated to outperform the state-of-art, with a fast enough computational cost for practical applications.

Overall, the focus of this thesis is the creation of a convolutional neural network based multimodal image fusion, specifically of brain images. Image fusion has been applied to many aspects that has been briefly introduced in section 2.3. In [30], a CNN is purposed for the fusion of different modalities, including CT, SPECT, MR among others. By the fusion via image pyramids, this paper aims to be consistent with human visual perception, that ends up leading to promising results in both visual quality and objective assessment. However, in this thesis we try to drop the mentioned image fusion strategy and instead define a differentiable perceptive loss function on a CNN architecture compatible with an image fusion problem. We train our CNN on preregistered publicly available medical image pairs of various grayscale modalities and obtain the final fused results solely based on our end to end learning based neural network.

3 Implementation

The framework proposed in this thesis is a convolutional neural network in order to get the image fusion. The pair of images that will be fused are MR-T2, CT among others. The first type i.e. MR-T2 is a specific time of MR, where the repetition time (time between the application of pulses applied on the desired slice) and the time to echo (the delay from the emission of the radio frequency energy and the reception of the echo signal) are both higher compared to the other most common MR imaging: the T1-weighted imaging. T1 ad T2 can be easily differentiated by looking at the CFS (cerebrospinal fluid), which appears as dark in T1 images and as bright in T2 [4].

The second imaging, i.e. CT, provides a faithful visualization of bones, soft tissues and blood vessels added to the capability of detecting tumors, swelling, bleeding, and tissue calcification. All these utilities and multi-detection added to its popularity and fast scan times has transformed CT to probably the most popular medical brain imaging technique [13]. [24] offers both an analysis and a comparison of MRI and CT in the particular field of study of stroke diagnosis.

The decided architecture used in this project has proved to show good results fusing the two mentioned modalities. There are two main reasons for using a convolutional neural network over using a normal fully connected neural network. The first reason is that this networks share a lot of parameters. Basically the meaning behind that is that a feature detector in a part of an image can be very useful too in another part of an image. Therefore, we can use the same parameters in different position of input images to detect both high and low level features. Another main reason is the sparsity of connections, where in every layer the output values only depend of a few inputs, converting the network into resilience to translation invariance. This property, also mentioned in section 3.3, means that the network is resistant to images that are slightly shifted.

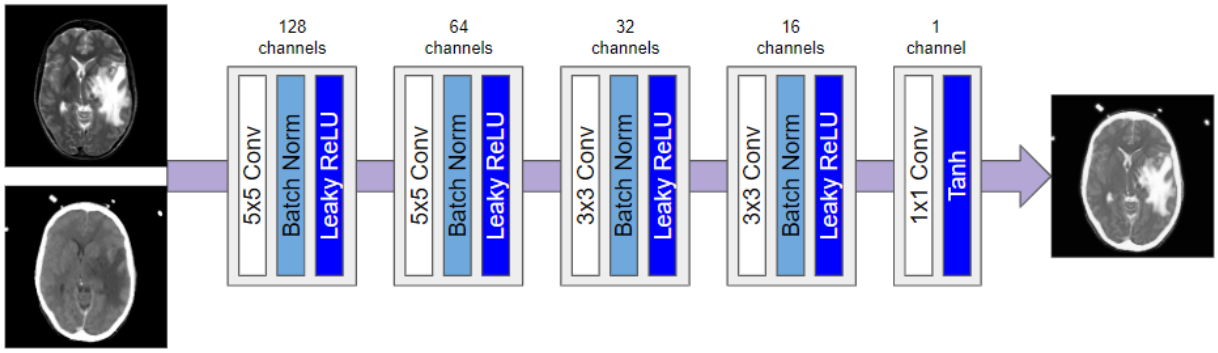

Figure 3.1: Architecture of the CNN.

Figure 3.1 shows the schematic representation of the end-to-end CNN architecture proposed in our work. As the starting point, we concatenate the two input grayscale images resulting in a two channel input. For the convolution operation, a 5x5 filter is used in the first and second layers, a 3x3 filter is used in the third and the fourth layers and 1x1 filter is used in the last layer. The striding in each layer is set to 1 and there is no padding operation performed during convolution resulting in no downsampling performed in our network. It is to be noted that every downsampling process will erase some detailed information in

the input images which is crucial for image fusion. We employ batch normalization and Leaky ReLU after the convolution operation in each of the first four layers in order to avoid the issue of vanishing gradient while the last layer has a tanh activation.

3.1 Preprocessing

First of all, the hole data base that will be used for the training has to be preprocessed. The images are all gray-scale images of 256x256 pixels. For training we have a total of 502 pairs of CT and MR images and all of them will be used for the training of the weights, biases and other hyperparameters. In some brain image fusion, not all the images are in grayscale as it can be seen in Figure 1.1. The results provided by this functions are very visible because the capability of mixing a RGB with a grayscale image creates a very attractive fusion, as the final image clearly has information from both grayscale and RGB images. However, the RGB images are usually coming from a false map, meaning that are obtained with monochromatic sensors and transformed into 3 channels for a more visual fusion. This conversion makes the computation harder and only adds distortion in order to get a better looking result.

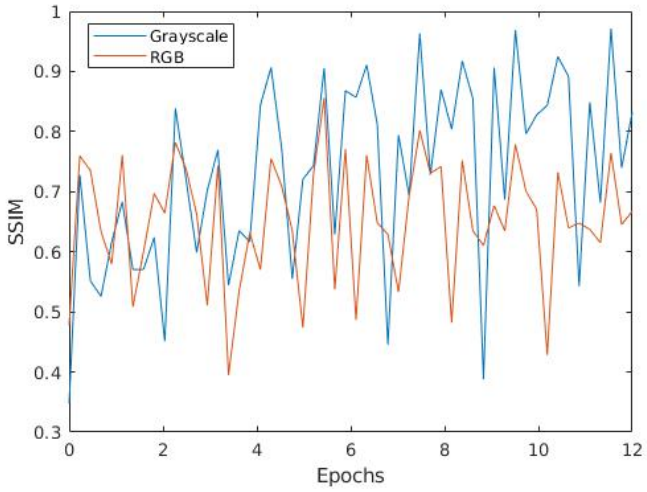

Figure 3.2: Loss function curve with respect to number of epochs on a combination of Grayscale and RGB based medical image fusion.

The Figure 3.2 shows the training with pairs of three channel images, where half of them are grayscale and the other ones are converted with a color map.As it can be seen along the training experiments, the network is optimizing the parameters for the grayscale part slightly but it unable to learn the RGB part at all and is rather inconsistent. We would like to focus our image fusion approach with a raw acquisition data i.e both grayscale inputs in order to attain a symmetry in our approach which will help to evaluate our method better. The only preprocessing that is going to be applied to the images is a normalization for the purpose that the entries are values between 0 and 1. The normalization it is not going to be between the maximum and the minimum of the images because this change disturbs the contrast of the image. Therefore, the only operation that is going to be applied is a division by the maximum of the dynamic range of the images, in this case of 8-bit digits we will divide or 255.

3.2 Network architecture

As in many other neural network problems, this network structure is not build from the scratch. Many networks previously designed have shown to have good results in similar problems to the ones that was initially designed for. The first layer tends to learn features that are not specific for a single assignment and consequently they are applicable to other datasets or tasks [44].

The architecture of the selected network is visible in Figure 3.1. The required entries are two two-dimensional 256x256 images that will be fed in the network as a tensor. The whole framework is going to be based on TensorFlow, an open DL library developed by Google. This system can be used to implement a wide variety of algorithms, including building and training algorithms for deep neural network model. This framework is widely used for many purposes and areas of work, including computer vision, text analysis, robotics or speech recognition among many others [6].

All the layers in the network follow a similar structure, specially the first four layers. The main two decisions that have to be taken are how is the convolution done and what kind of activation layer will be used. In this section the decisions involving the convolution and the activations will be explained and justified.

There are different options for the initialization of the weights that are going to be learned in the network. In the Figure 3.3 we can see the training loss evolution of a basic convolutional network. The network learns to classify handwritten digits obtained from 60000 digit scans from the MNIST dataset [2]. The graph shows the training results for three identical networks with different types of initialization of the weights of the network.

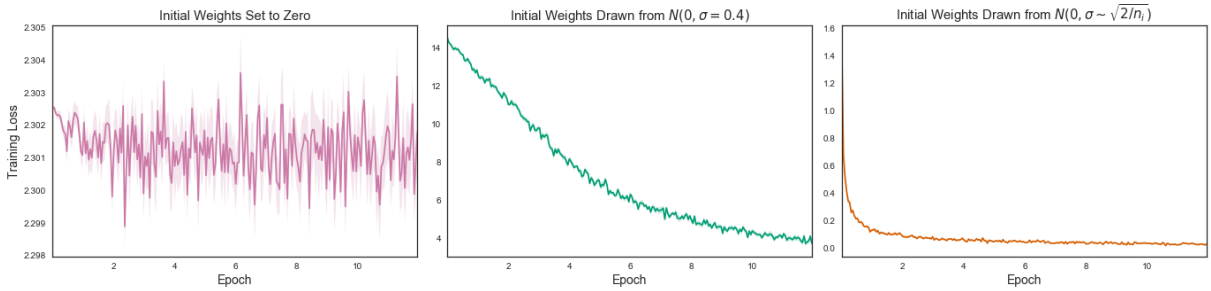

Figure 3.3: Loss function evolution in 12 epochs from [3].

The importance of the initialization shines in the graphs of Figure 3.3, where with thee different types we get complete different training results. The first case is initialization of all the weights to 0, which basically transforms the model into a linear one. Therefore, the network can not learn and the result is arbitrary for every iteration and we get a very noisy output. The second image selects the weights from a normal distribution with an arbitrary standard deviation of 0.4. Now the network is able to correctly learn after each epoch but the convergence to a good result is very slow. The accuracy achieved in this case after the 12 training epochs is about 88%. The third case is the usage of $\sqrt{2/n_i}$ as the variance of the normal distribution. The usage of this variance is found to have state-of-art results when the activation layer is ReLU, and was proposed and demonstrated in [20].

In our case, we will use a standard deviation of 0.001 that prove to have a good behaviour and stabilizes the loss function in few epochs. We will more specially use a truncated normal distribution. The only difference between a normal and a truncated normal is that in the second case values that differ more than two standard deviations of the mean of the distribution will be discarded and redrawn. This way we

assure that all values will be random small values close to 0 as it was desired.

Not only the weights have to be initialized but also the biases. The bias is a kind of threshold for each neuron that determines whether it is activated or not. All the biases have to be then initialized and trained. A very common initialization is to set all the biases to 0 because in this case the network is going to single-handedly learn an appropriate value for each neuron. In addition, if any bias is not required or used it is already initialized to 0 and it won't be used.

To execute the convolution we will use a stride of one. The reason we do that is that the final output of the network has to be of the same size that the true entries so that the fusion makes sense. Therefore we will always keep the same dimensions during the whole convolution. We will also use padding in order to maintain this image size.

After the convolution we will use batch normalization [21]. This kind of normalization gives a solution to the phenomenon known as covariance shift. TUsing this normalization, the parameters only are modified with the information of the change of the parameters of the previous layer. This makes training more costly and difficult because it requires low learning rate and very careful and accurate parameter initialization. With batch normalization we are able to use higher learning rates which accelerates the learning process that has lead to a great growth of the popularity of algorithms based in neural networks.

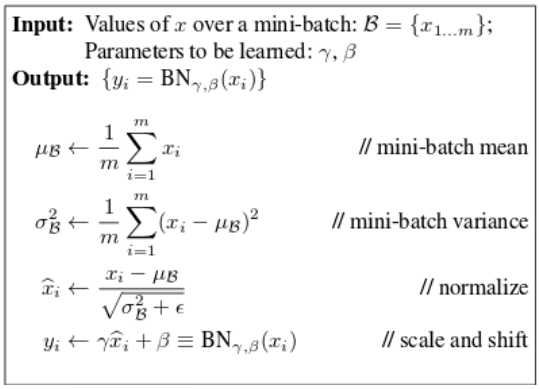

Figure 3.4: Batch normalization from the original paper [21].

The Figure 3.4 shows the batch normalization algorithm. The idea behind it is to normalize the inputs to be 0 mean and of unitary variance. The addition of the batch normalization is to add the parameters β and γ so that the mean and variance values can change to whatever the network requires in every node. By adding this two parameters, each layer does the training in more stable conditions and in a more accelerated learning process. We will also sum a small parameter ϵ to the variance in the normalization in order to assure that the output is never divided by 0 and making it this way even more stable. With this sum we also provide some strength against overfitting.

After the convolution and normalization what has to be applied is a non-linearity function. This function is the one that makes the network viable and is the one that enables the learning of the features. This layer saturates the output in order to either adjust or cut-off the values. Some standard models use the sigmoid $f(x) = \frac{1}{1+e^{-x}}$ or the tangent $f(x) = tanh(x)$ [27]. This two functions take higher computational times compared to the ReLU activation function [17]. Even some other easier functions might come to mind like a binary step function. However, we require a function whose derivative is not constant because in order to train the network we are computing the gradient of the function that are send for the calculation of the errors. If the gradient is 0 the network is not correctly trained since there is no improvement hap-

pening. This functions are sometimes used but for other purposes that will be explained at the end of this section.

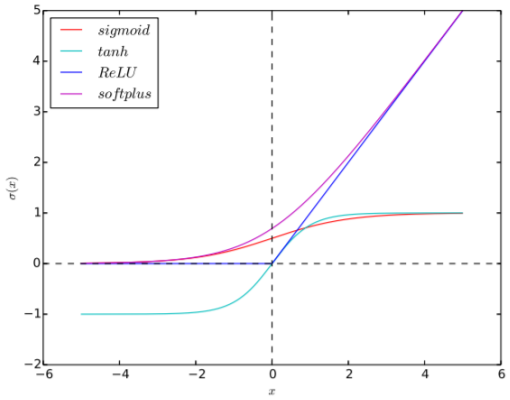

Figure 3.5: Popular nonlinear activation functions [10].

The sigmoid function is clearly nonlinear and has continuous derivative and we can correctly use backpropagation and the network will be able to be trained. The function clearly saturates the values that are far from the point 0. The problem is that when the values are not close to this middle point the function is almost flat and therefore the gradient values are very small. That means that for extreme values the network is not going to learn appropriately.

Another problem that the sigmoid function has is that all the output values are limited from 0 to 1. Consequently, the function is not inverse symmetric and only gives positives values which is not always desired. To solve this problem the $tanh$ function started gaining usage, which is a escalated and shifted version of the sigmoid function ($tanh(x) = \frac{2}{1+e^{(-2x)}} - 1$) ranging from -1 to 1. With this change we are able to solve the negative values problem but we have the same vanishing gradient.

Then the ReLU function was proposed: $f(x) = max(0, x)$. As it can be seen in the Fig. 3.5, it is easy to check that it is a nonlinear function where we can propagate the errors. The main advantage of this function is that if we get a negative input we will get a null output so that neuron is not activated. With this fact, not all the neurons have to be active at the same time making it very efficient for computation. The gradient of the function is the binary step one mentioned previously. With this, the negative input values have zero gradient and are not updated during backpropagation that can create dead neurons.

A way to solve this problem was proposed as the activation function that we will mainly use, the Leaky ReLU [44]. The innovation of leaky ReLU is that it adds a small slope to the negative part. With this addition we are not finding the zero gradient problem anymore and we will not get any dead neurons anymore.

As it can be seen in Fig. 3.1, the only layer that is a bit different and has a different activation layer is the last one. The main reason for that is that the output we need has to be the fused image and consequently the values have to be in the range between 0 and 1. The intuitive solution is then that the activation layer should be the sigmoid which already is bounded in that limits but the one used is going to be tanh.

The reason behind it is the compute of the gradient for the backpropagation, because the derivative of the tanh function can be calculated with the original function: $\frac{\partial tanh(x)}{\partial x} = 1 - tan^2(x)$. This fact makes this

operation computationally faster and during the learning, the network itself will keep the output values between 0 and 1 in order to get the best results.

3.3 Training

In order to let our network train, we have to define a loss function to evaluate the outputs and let the network learn how is it performing so that the network can improve and provide better results. We have to compute the gradient of this loss so that the network learns how to change the parameters to obtain a lower cost function. The process of calculating this gradient and pass it to previous layers to train the network is known as backpropagation. This algorithm is probably the most widely used among all the supervised learning algorithms [11].

For every entry of information to the network, all the convolutions and nonlinear operations are executed and a loss function is computed with the output that we got. Then we go back layer by layer computing the gradient of this loss function in order to improve it in the next step until we get to the inputs and we have recomputed the weights and biases. This forward and backward propagation is executed for every batch in the training and repeated for each desired epoch.

We have to then choose what is the desired loss function that is going to evaluate the performance of the network and determine how the training is going to develop. As an evaluation metric, we would like to get a subjective evaluation of the fusion according to the Human Visual System (HVS). The only possible way to do it completely adapted to HVS would be to have some trained observers to score the fusions, which requires the human intervention and high cost, time consuming and unrepetible [36]. Therefore, it is required to have a metric that combines both objective and subjective evaluation taking into account the HVS and knowing that standard quality image metrics like PSNR or MSE do not give an accurate evaluation of the fusion performance.

In [42], Structural Similarity Index Measure (SSIM) is presented. This metric is faithful with HVS and its goal is to improve the quality assessment by focusing the attention on structural information, which is very sensitive to humans visual system. This index is based on the product of three different components that range from 0 to 1: saturation, luminance and contrast comparison.

$$SSIM = l(x,y)^{\alpha} * c(x,y)^{\beta} * s(x,y)^{\gamma} \tag{3.1}$$

where $\alpha > 0$, $\beta > 0$ $\gamma > 0$ are parameters for adjusting the importance of each component. Before decomposing this equation to all three comparison components, we define the estimation of the mean, the variance and the correlation coefficient like in the following equation.

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma_x = (\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2)^{\frac{1}{2}} \tag{3.2}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{3.3}$$

We can define the luminance comparison as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{3.4}$$

Where $C1$ is a constant included to avoid instability when the means are close to 0. In the paper [42] the constant is defined as $C1 = (K_1 L)^2$ where L is the maximum of the dynamic range of the pixels and $K_1 \ll 1$ is a small constant and since we are using 8-bit representation for our images $L = 255$. The contrast comparison component is defined as:

$$c(x, y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3.5}$$

Again a similar constant $C2$ is added for the same purpose with the same definition $C2 = (K_2 L)^2$. Last but not least, the definition of the structure or saturation comparison is:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{3.6}$$

This third constant has the same properties than $C1$ and $C2$ but in the original paper it is defined as $C3 = C2/2$. All three components verify three conditions required for the compute of the metric:

1. They must be symmetric: $S(x, y) = S(y, x)$
2. It must be bounded with a maximum of 1: $S(x, y) \leq 1$
3. It must have a unique maximum: $S(x, y) = 1$ if and only if $x = y$

The HVS sensitivity to luminance change depends on the background luminance, not only the absolute luminance. This phenomena is known as luminance masking. For our loss function, we need to have a metric that is consistent with the human visual system so we get at the end a suitable for human evaluation fused image.

In order to visually improve the results, we would desire more importance of the luminance component over the other two. Getting back to (3.1), the standard version of SSIM uses $\alpha = \beta = \gamma = 1$. The values can be changed if the goal is to give more importance to one of the three components. In the case of our database we clearly want to give α a higher value. The three mentioned conditions are still being accomplished with the change in one of the three parameters.

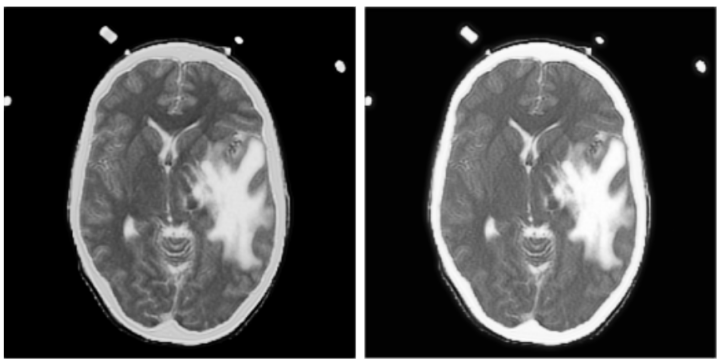

Figure 3.6: Fusion after training with $\alpha = 1$ (left) and with $\alpha = 4$ (right).

In Fig. 3.6, we see the comparison of the two images using different alpha values. It is visible that the final fused image is brighter and gives a better looking result. Even the contrasted points are a bit blurred

and we have a lose of information in the textures, the fusion giving higher α values are visually more pleasant. By giving more importance to luminance we get better results that will be discussed in the next section, where $\alpha = 4$ is going to be used.

We also need to pick up an optimizer for our values. A very commonly used optimizer for neural networks is the gradient descent optimizer. It consist on updating the parameters of the network in the opposite direction of the gradient of the loss function. It requires the decision of the learning rate (LR) parameter, which is the size of the step that is going to be done in each iteration to reach a minimum of the function. The selection of the learning rate is not trivial because a too big LR might not lead to the most optimal point by fluctuating around it or even diverge, but a too slow rate can lead to very slow convergence to the appropriate result.

For our network we are going to use the Adam optimizer [26]. This optimizer uses moving average of parameters known as momentum to compute more effective learning rates to converge to the solution faster. This way, the given learning rate to the optimizer is not that important. The only small modification we have to apply is that this optimizer tries to minimize the input parameter. As SSIM is a metric of similarity we need to maximize instead of minimizing. An easy way to transform into a maximization the input of the optimizer has to be the SSIM obtained but changing the sign.

4 Evaluation

In this section, all the results will be introduced. But before getting to the fusion we must define how are we going to evaluate the fusion, which is a big uncertain and of multiple choices. We will first introduce the two metrics that are used for the evaluation and then we will present the other fusion methods that will be used to compare the results.

4.1 Evaluation metrics

For the evaluation of the fusion we will compare with two different metrics. The first one is going to be the already presented SSIM. The exact metric that we will be using is a modification of SSIM but it is also based in structural similarity, where the loss of this structural information can be a good approximation of the perceived distortion. This metric was first introduce in [34] and it has an open source code [35].

The main reason behind the use of this metric and not the normal SSIM is that the metric measures how good the fusion is based on both the original image and the fused one, when in the normal SSIM, the evaluation is between only two images.

The other metric that is going to be used is the feature mutual information, introduced in [19]. This is a non-reference metric which calculates the information preserved in the fused image to give an evaluation to the final image. The paper shows the performance of the metric in diverse data sets and it proves the efficiency and higher consistency with subjective evaluation.

4.2 Compared methods

In this section some other fusion methods will be introduced. This methods are the ones that are going to be compared with the proposed fusion algorithm. The methods are weighted averaging, guided filtering and another method based in CNNs.

4.2.1 Weighted averaging

This method is probably the most simple method function. The fusion is going to be based on giving an importance value to one of the images and another one to the other. The values have to sum up to 1 in order to not create a lighting distortion. Since this is a very simple method we will give the same importance to both the input images. A simple representation of the structure can be seen in the Figure 4.1.

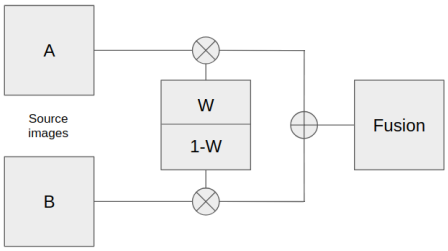

Figure 4.1: Architecture of weighted averaging algorithm.

4.2.2 Guided filtering

This is an image fusion algorithm purposed in [29]. This algorithm (that can be seen in the next figure) is basically based in the decomposition of the images to be fused in one that contains large scale intensity variations and another one containing the details. The way to obtain the first decomposition is by applying an average filter to the entire image (in the case of the paper a 31x31 filter). After getting the average decomposition the way to obtain the detail one is by subtracting the first one with the original.

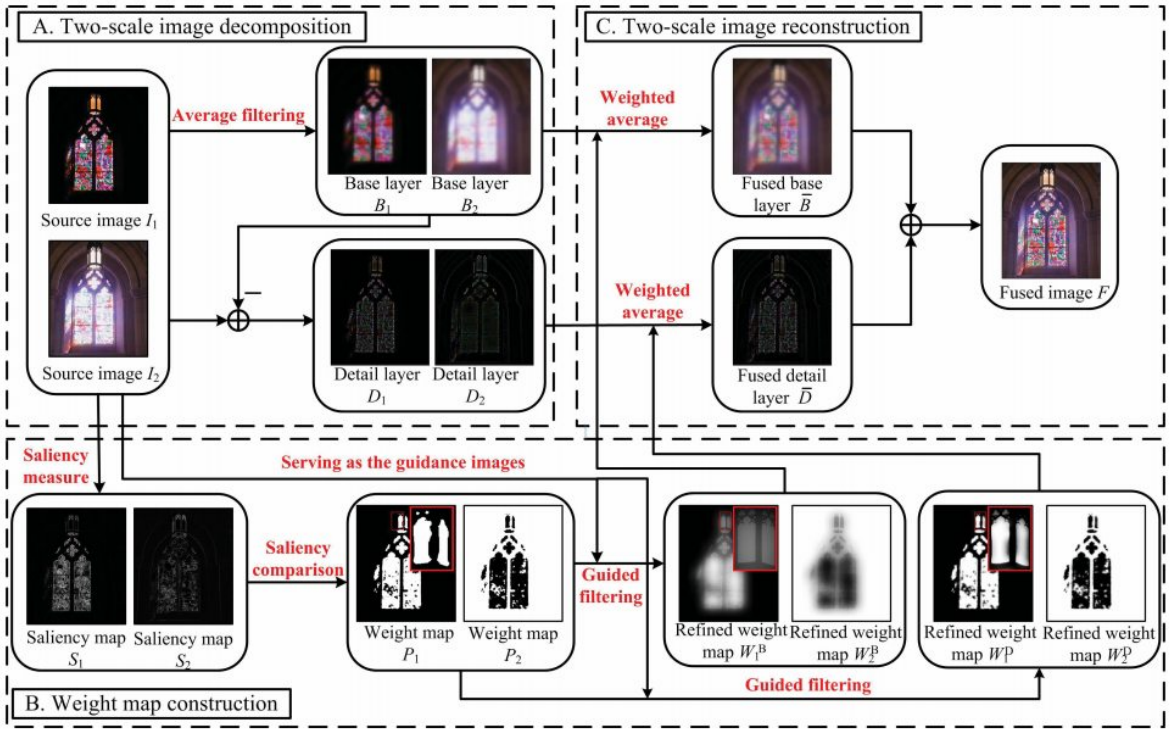

Figure 4.2: Guided filtering fusion algorithm procedure proposed in [29].

By applying some guided filtering and comparisons, a weighting map is obtained which shows what part of each images have more or less relevance in the final fused image. This map is applied by weighted averaging to the images and the resulting image is obtained. The complete structure of the guided filtering scheme can be seen in Figure 4.2.

4.2.3 Laplacian Pyramid-Convolutional Neural Network (LP-CNN)

This third method is introduced in [30]. This is a different approach than the proposed in this thesis, where we directly obtain the fused image at the end of the network. In this paper, the method is implemented to get the weighting map at the output of the CNN. Since the output of the network is not the fused image, the architecture of the network is really different, and all the decision explained are not necessary the same.

The fusion is done a bit different than in the other explained methods. The images are decomposed in Laplacian pyramids. Finally, the used coefficients are fused with the help of the weighting coming out of the CNN. The fusion algorithm can be seen in the following figure.

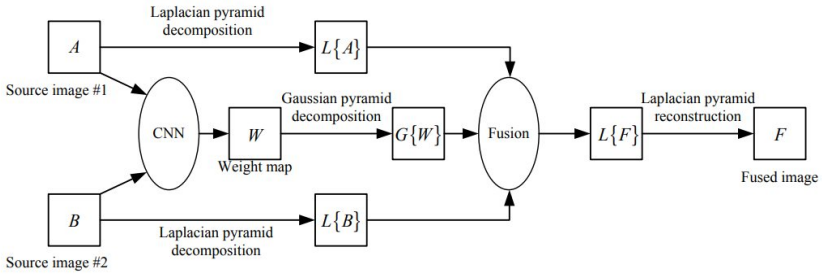

Figure 4.3: Fusion diagram obtained from [30].

4.3 Results

Before presenting the results of the fusions, Figure 4.4 shows how the loss function evolves during the training. As it can be seen, the network is clearly able to improve the fusion of both MRI and CT images as the loss function approaches correctly the desired minimum. For the last few epochs the loss function does not increase that much but the training evolves definitely better that the one with three channel images (Figure 3.2).

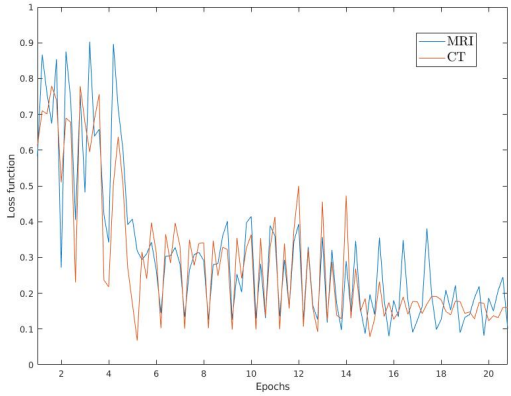

Figure 4.4: Evolution of the loss function during the training of the network.

For the testing we will only use two random pairs of images. The following results are the evaluation metrics obtained with the three methods and the proposed one.

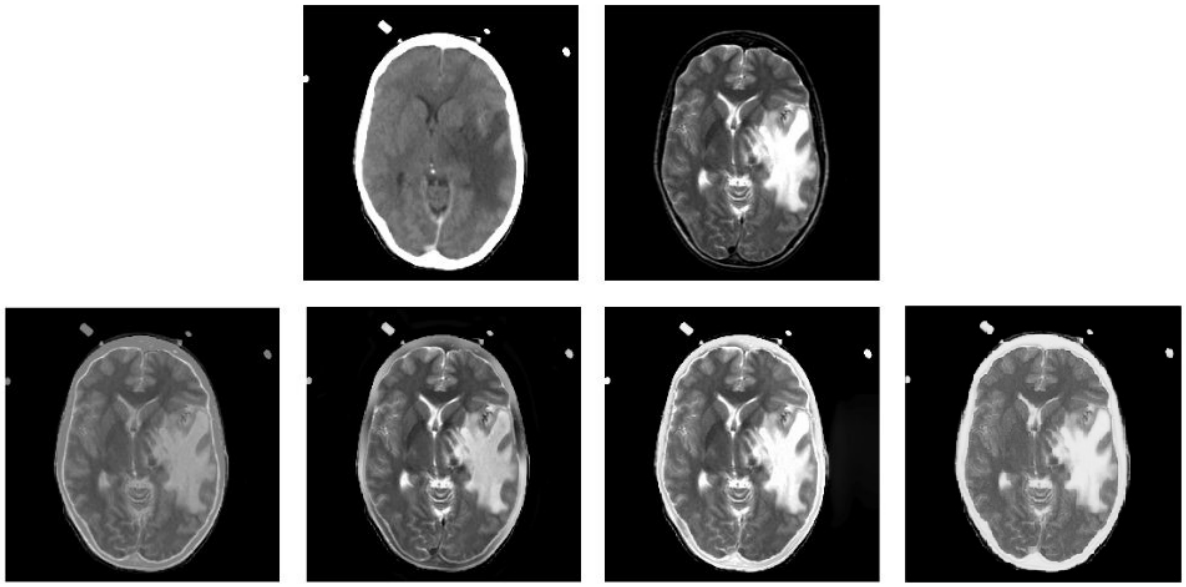

Figure 4.5: The two top images are the image to be fused for the first test: top-left CT, top-right MRI. At the bottom from left to right: weighted averaging, guided filtering, LP-CNN and the proposed method.

	WA	GF	LP-CNN	Proposed
FMI	0.8926	**0.9043**	0.9016	0.8967
SSIM	0.8382	0.8393	0.8625	**0.8659**

Table 4.1: Results of the first fusion.

As it can be seen in the Table 4.1 there is a lot of discordance on how should a fusion be evaluated. We can see that with the SSIM metric the proposed method has the better performance but with the FMI metric, guided filtering gets better results.

It can be clearly seen that the weighting average method performs poorly compared to the other three methods but it still is quite close. It can be seen that LP-CNN does not have the better performance looking at any metric but still it can not be said that is worst than the other two methods because LP-CNN outperforms each fusion depending on the evaluation method.

In Figure 4.4 the images are presented. As it can be seen, the results are very diverse and each method performs well in different aspects. Taking a look to the Figure 3.6, the result might be more visually attractive than the one in this last figure but this one gets way higher performance results.

In both guided filtering and the CNN based in Laplace decomposition, we seem to get attractive results because features seem to be more visible. However, some fake shadows appear and the textures of the CT original image are not that visible. The weighted averaging and the proposed method get less brightening results that are not as visually pleasant. It is clear that the weighting average only gets very gray results for the part that both images are very different instead of showing the features of one of the two original images.

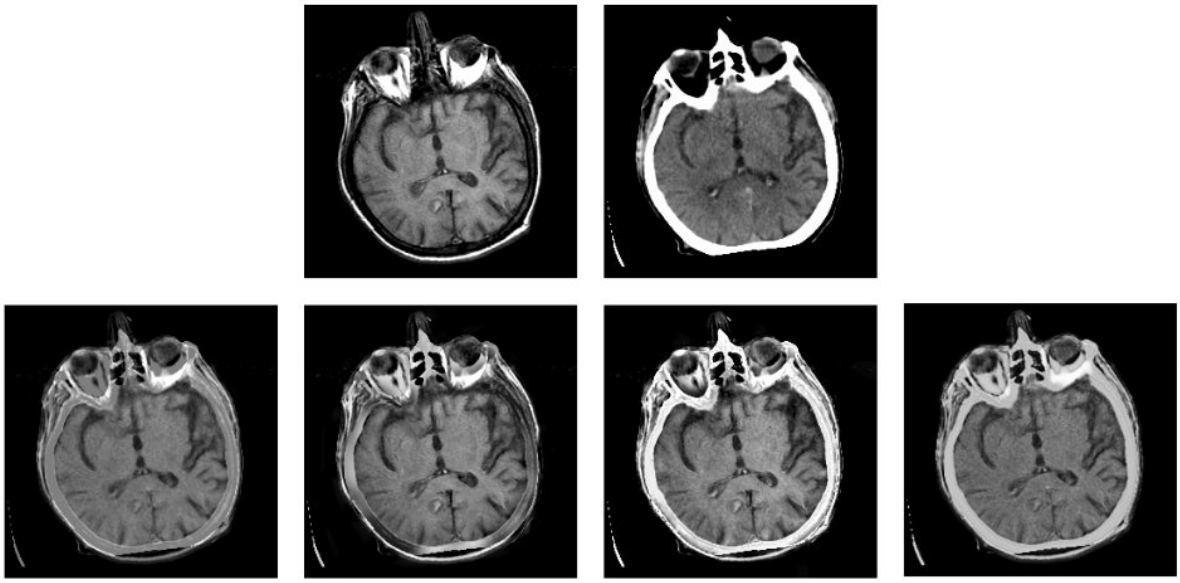

Figure 4.6: The two top images are the image to be fused for the second test: top-left CT, top-right MRI. At the bottom from left to right: weighted averaging, guided filtering, LP-CNN and the proposed method.

	WA	GF	LP-CNN	Proposed
FMI	0.8716	**0.8860**	0.8801	0.8731
SSIM	0.7892	0.7991	0.7936	**0.8067**

Table 4.2: Results of the second test fusion.

An important aspect that can be clearly seen in guided filtering and not as much in LP-CNN is that they give a very strange fusion for the cerebrospinal fluid, which changes of color around the whole brain. Even this fact seems to be very visually less attractive this methods still get very good performances. This problem clearly makes the problem of a lack of objective fusion evaluation shine, since the metrics are still not able to totally define what is better or worse.

We can see for the second table that we obtain very similar results between all the methods, being the proposed one the outperforming according to SSIM and again guided filtering works better with the FMI metric. We can also see the appearing of the shadows in both guided filtering and LP-CNN but they also get a less gray result and is maybe visually more attractive.

5 Conclusions

As it can be seen in the results we were able to successfully build a network that is able to fuse two different modalities medical images. We fused pairs of 256x256 gray scale images without the help of any fusion rule or decomposition, just by feeding in the raw pixels to the network.

The main innovation of this project is this lack of fusion rule or decomposition, since other methods required a previous and later step in order to get the final output. It also has the advantage of being a end to end learning based algorithm, since every layer of the network as well as the loss function are differentiable. This fact shows the great potential of this method and the big possibility of improvement.

The results show that the network is trained to maximize the SSIM and not the FMI. This happens because the optimization of the parameters only care about the first metric and not the second one. A possible good improvement is to train the network with more than one loss function, or in other words use a loss function that is combination of more than one metric.

One of the main obstacles found during the thesis was the low computational capacity of the used computer during the developing of the network. The construction of the final network that was used to get the presented results took more than 8 hours to be trained, with a single GEFORCE GTX 1050 Ti GPU. The lack of access to a better computational unit did not allow the best possible timings although its possible that our network learns the best value of its parameters within seconds given multiple GPUs.

To sum up, a new algorithm for image fusion was developed with high quality outputs and state-of-art results. Even the results are not the best for all the metrics, it shows really impressive performance in all of them getting the better result in a reliable image quality metric.

Bibliography

[1] http://www.med.harvard.edu/aanlib/.

[2] Mnist dataset: http://yann.lecun.com/exdb/mnist/.

[3] Mnist graph: https://intoli.com/blog/neural-network-initialization/.

[4] Mri basics: http://casemed.case.edu/clerkships/neurology/web

[5] Web of science. http://www.webofknowledge.com.

[6] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

[7] T. Akbarpour, M. Shamsi, and S. Daneshvar. Medical image fusion using discrete wavelet transform and lifting scheme. In *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pages 293–298, Nov 2015.

[8] Robert J Amdur, David Gladstone, Kenneth A Leopold, and Robert D Harris. Prostate seed implant quality assessment using mr and ct image fusion. *International Journal of Radiation Oncology*Biology*Physics*, 43(1):67 – 72, 1999.

[9] V. Barra and J. . Boire. Quantification of brain tissue volumes using mr/mr fusion. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, volume 2, pages 1451–1454 vol.2, July 2000.

[10] Hoon Chung, Sung Lee, and Jeon Park. Deep neural network using trainable activation functions. pages 348–352, 07 2016.

[11] Yann Le Cun. A theoretical framework for back-propagation, 1988.

[12] D. Dey, D. G. Gobbi, P. J. Slomka, K. J. M. Surry, and T. M. Peters. Automatic fusion of freehand endoscopic brain images to three-dimensional surfaces: creating stereoscopic panoramas. *IEEE Transactions on Medical Imaging*, 21(1):23–30, Jan 2002.

[13] W. M. Diyana, W. Zaki, and CunRui Kong. Identifying abnormalities in computed tomography brain images using symmetrical features. In *2009 International Conference on Electrical Engineering and Informatics*, volume 01, pages 88–92, Aug 2009.

[14] Su Liao Qingmin Bloyet Daniel Constans Jean-Marc Chen Yanping Dou, Weibei Ruan. Fuzzy information fusion scheme used to segment brain tumor from mr images. pages 208–215, 2003.

[15] El-Sayed Rabaie Abd Elrahman Wael Allah Osama Abd El-Samie Fathi. Elhoseny, Heba El-Rabaie. Medical image fusion: A literature review present solutions and future directions. 2017.

[16] El-Sayed Rabaie Abd Elrahman Wael Allah Osama Abd El-Samie Fathi. Elhoseny, Heba El-Rabaie. Medical image fusion: A literature review present solutions and future directions. 2017.

[17] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[18] R. J. T. Gorniak, E. L. Kramer, G. Q. Maguire, M. E. Noz, C. J. Schettino, and M. P. Zeleznik. Evaluation of a semiautomatic 3d fusion technique applied to molecular imaging and mri brain/frame volume data sets. *Journal of Medical Systems*, 27(2):141–156, Apr 2003.

[19] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. A non-reference image fusion metric based on mutual information of image features. *Computers Electrical Engineering*, 37(5):744 – 756, 2011. Special Issue on Image Processing.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[22] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4 – 19, 2014. Special Issue on Information Fusion in Medical Image Computing and Systems.

[23] Filip Sroubek Jan Flusser and Barbara Zitov a. Image fusion: Principles, methods, and applications. 2007.

[24] R. S. Jeena and S. Kumar. A comparative analysis of mri and ct brain images for stroke diagnosis. In *2013 Annual International Conference on Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy*, pages 1–5, June 2013.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[28] H. Li, R. Deklerck, B. De Cuyper, E. Nyssen, J. Cornelis, and A. Hermanus. Object recognition in brain ct-scans: Knowledge-based fusion of data from multiple feature extractors. *IEEE Transactions on Medical Imaging*, 14(2):212–229, 1995. cited By 44.

[29] S. Li, X. Kang, and J. Hu. Image fusion with guided filtering. *IEEE Transactions on Image Processing*, 22(7):2864–2875, July 2013.

[30] Y. Liu, X. Chen, J. Cheng, and H. Peng. A medical image fusion method based on convolutional neural networks. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–7, July 2017.

[31] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Processing Letters*, 23(12):1882–1886, Dec 2016.

[32] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191 – 207, 2017.

[33] Yu Liu, Xun Chen, Zengfu Wang, Z. Jane Wang, Rabab K. Ward, and Xuesong Wang. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Information Fusion*, 42:158 – 173, 2018.

[34] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):94–109, Jan 2012.

[35] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):94–109, Jan 2012.

[36] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153 – 178, 2019.

[37] M. Mysore, Nandeesh Meenakshi. Image fusion algorithms for medical images-a comparison. 2015.

[38] Rahul P. Mundhe Juilee M.Ghatole Prof. Anuradha S. Deshpande, Dhanesh D. Lokhande. 4, 03 2015.

[39] Kusum Rani and Reecha Sharma. Study of different image fusion algorithm. 9001, 04 2008.

[40] Mansoor Raza, Iqbal Gondal, David Green, and Ross L. Coppel. Classifier fusion to predict breast cancer tumors based on microarray gene expression data. In Rajiv Khosla, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, pages 866–874, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[41] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[42] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[43] Yunfeng Wu, Cong Wang, S. C. Ng, Anant Madabhushi, and Yixin Zhong. Breast cancer diagnosis using neural-based linear fusion strategies. In Irwin King, Jun Wang, Lai-Wan Chan, and DeLiang Wang, editors, *Neural Information Processing*, pages 165–175, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[44] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[45] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. *CoRR*, abs/1504.03641, 2015.

[46] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, June 2010.

Acknowledgments

I would first like to thank Dr. Prof. Stefan Gumhold for providing me the topic that better fitted my priorities and knowledge, added to some guidance and problem solving during the thesis. I want to give an special thank to my home university professor Dr. Javier Ruiz-Hidalgo, for facilitating the abroad development of the thesis and helping with the communication of both universities. Finally, I want to express my very profound gratitude to MSc. Nishant Kumar, for being my big help and guidance during the whole development. Without his advice, support or willingness to help I wouldn't have been able to complete the thesis.