Universität Duisburg-Essen
Department of Computer Science and Applied Cognitive Science

Seminar Report
Master Computer Engineering(ISE)

Machine Learning

Regularization of CNN using Shape Prior

Neelu Madan

Date: 26. June 2018

Abstract

A goal in machine learning is to create models that generalizes well to new data. Different regularization strategies to improve the generalization capability of a neural network model are used. These regularization strategies could be general or specific to the task. This report focuses on regularization of CNN using shape prior, which is mostly used for bio-medical image processing. The idea is to incorporate the information about the shape and location of organ in human body while training neural networks. As a result, model can predict the disparity or irregularities in organs more accurately.

Contents

1 Introduction

1.1 Motivation

Most classification and the regression models utilize a pixel level loss function such as cross-entropy or mean square error, which does not completely consider the hidden semantic data and overall contingency in the output space. For example, we need the semantic or global information while predicting the class labels in an image. This additional semantic information could be incorporated by training neural networks with additional "regularization" term. For instance, this regularization term might ensure the smoothness of object boundaries, which gives cues about the geometry of the object. One could vary this regularization term being used as per the requirements. In this report, we added "shape priors" as a regularization term for medical image processing networks. The following approaches will be examined:

1. Anatomically constrained neural networks (ACNN): It is a principal component analysis (PCA) based approach.

2. GridNet: Based on registration of shape prior on input image.

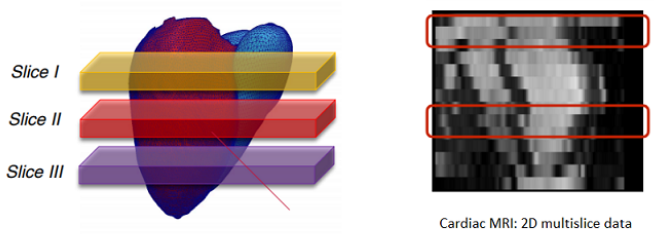

Figure 1.1: Motion artifact in acquisition of the MRI image.

1.2 Problem Statement

Conventional image segmentation and super resolution networks are not performing well over the medical images because of motion artifacts as shown in [**fig:1.1**]. Image acquisition using MRI is a slow process and individual's motion (for e.g. systole and diastole for heart) makes it even more difficult to capture the stable images, this lead to the motion artifact. It is difficult to perform analysis on such bad quality data. We already have the information about the location and shape of the organ in the human body from the human anatomy. Therefore, we could supply the neural networks with

this additional information about the shape of the organs, also called as "Shape Prior". This information will further used to regularize the neural network models and make the predictions that confirms the shape of the organ in human body. As a result, the respective segmentation or super resolution network could remain unaffected by the large amounts of background clutter, noise and partial/complete occlusions.

This prior information could take multiple forms, especially for bio-medical image processing. A few of them are mentioned below:

- Boundaries and edge polarity [1]
- Shape models [2]
- Topology specification
- Distance prior between regions
- Atlas priors [3] presented by Bai et. al.: Appropriate for medicinal imaging applications since they authorize both location and shape priors.

We could also add above information in addition to the "shape prior" regularization to further improve the accuracy of the model.

1.3 Outline

In chapter 2, different regularization techniques used in neural networks will be explained in detail. Chapter 3 contains the information about auto-encoders and T-L embedding network used for the regularization of ACNN (Anatomically Constrained Neural Networks). In chapter 4 and chapter 5, details of the neural network with new objective function to incorporate the prior knowledge about the shape will be elaborated. chapter 6 concludes the report with brief revisit of regularized network and conclusion.

2 Regularization

Regularization can be motivated as "a technique to improve the generalization ability of a learned model". [4]. To elaborate it further, the major issue with the machine learning model is that the models are performing well on training data. However, the same models are failing to give the desired performance on new data. There are some pre-existing techniques to reduce the error on test data at the expense of increasing error on training data. These techniques are called as the "regularization" and may differ as per the input training data. There are lots of other factor, which may influence our choice of regularization techniques such as amount of data, type of data etc.

For deep learning models we need huge amount of data, which further helps to increase the generalization capabilities of model. With increases in training data, the training parameters also increases to approximate the good representation of data. This result into over-fitting of deep networks to the training data. Consequently, the model become overly complex and representing only training data. Such models are not susceptible to noise or the outliers in this case. As the incoming data is unpredictable, therefore we are not sure about the estimated model whether it'll be performing well on real world data or not.

Furthermore, we need a way to control this over-training on data and at the same time increase the generalization capability, is also well know by a short term called "Bias-variance dilemma". The aim is to find an optimal balance between the bias (over-fitting) and the variance (under-fitting) for a neural network model. However, it is hard to find that stability between the bias and the variance, as whenever we try to decrease the bias (generalize the model) the variance automatically increases (over-fitting) and vice-verse. To achieve this, we add an extra "regularization term" in the training objective. A good regularization finds the optimum balance between bias and variance.

The major challenge is now to tune this regularization term, which is quite tedious task. As already mentioned above, as we try to decrease the over-fitting in neural networks, the under-fitting automatically increases and vice verse. A good machine learning model tends to minimize both under-fitting and over-fitting, but both can't not be low at the same time. There is no upper hand, but usually under-fitted models are preferred over the over-fitted model. This is due to the fact that a under-fitted models are more generic. The most common techniques to achieve regularization in deep learning models are cross-validation, dropout, data augmentation, L1/L2 regu-

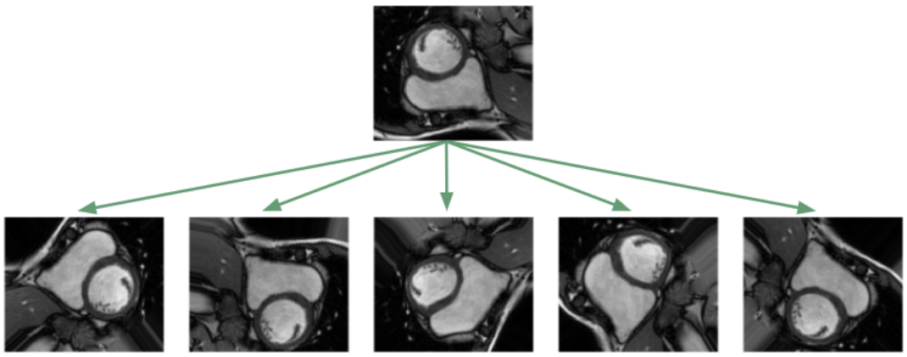

Figure 2.1: Data augmentation for cardiac MRI images

larization etc.

Two most widely used techniques have been described in this chapter:

- **Data Augmentation**: A special type of regularization used mostly for medical image segmentation, where training data is very small. Therefore, training samples are been increased by applying various transformations to the existing data.
- **L1/L2 Regularization**: A very common type of regularization used in almost all the neural networks, that penalizes large weights. The idea is to prevent over-fitting by avoiding over-learning of any features.

2.1 Data Augmentation

Due to unavailability of sufficient amount of training data, machine learning models are unable to generalize well to real world data. One solution to this problem is simply increase the training data. The new extended data-set could be generated from existing data by transforming existing data-set and adding it to the training set. Such transformation for images includes rotation, scaling , shearing etc.

Data augmentation is well suited for classification because the task of the classifier is to convert a high dimensional input (an image) to a single class. Thus a stable classifier needs to be invariant to all kind of transformations. For example, classifier trained without additional augmented data will predict different classes for the same ball as shown in [**fig:2.2**]. However, this approach may not be well suited for many other types of machine learning task such as generating new fake data for a density estimation. This is due to the reason that, we need to solve the density estimation problem to generate the fake data. Additionally, solving a density estimation is tedious task especially when we don't have enough samples. The estimation changes every time with increasing sample size.

Deep learning tasks usually requires a huge amount of diverse data. The model will not be really general when we train it on small data-set. By this techniques we are not only increasing the amount of training data, but also improving the generalization capability of the model. Sometimes, a few transformation might change the correct class. Therefore, one should be very careful while applying those transformation. For example, optical character recognition (OCR) tasks requires to distinguish between "b" and "d" and the difference between "6" and "9," so horizontal flips and 180 degree rotations are not proper methods for increasing data-set for these undertakings. We could enforce an additional penalty factor for the such transformations.

Figure 2.2: Tennis ball with certain translations

2.2 L1/L2 Regularization

L1/L2 regularization is the most common type of regularization technique used for most machine learning or deep learning algorithm. Usually the two decisions are:
1) L1-norm vs L2-norm loss function
2) L1-regularization vs L2-regularization.

The error function for L1 and L2 norm are mentioned below:
L1-norm loss function: It is also know as least absolute deviations (LAD) or least absolute errors (LAE). It minimizes the sum of the absolute differences (D) between the real worth (Vi) and the estimated worth (f(ui)):

$$\mathrm{D} = \sum_{i=1}^{n} |v_i - f(u_i)|$$

L2-norm loss function: It is also known as least squares error (LSE). It is basically minimizing the sum of the square of the differences (D) between the real worth (Vi) and the estimated worth (f(ui):

$$\mathrm{D} = \sum_{i=1}^{n} (v_i - f(u_i))^2$$

It is difficult to decide which one of the two L1 or L2 should be preferred. The usability is generally depends upon the data-set or requirement of the machine

learning algorithm. For example, L1 regularization could be used for feature selection and L2 regularization could be used for suppressing weights to prevent over-fitting. The differences of L1-norm and L2-norm as a loss function can be quickly summarized as below:

L2 loss function	L1 loss function
Not very robust	Robust
Stable solution	Unstable solution
Always one solution	Possibly multiple solution

2.3 Regularization using shape priors

In this regularization technique, a few logic shape representation of organs in human body is given by the experts. This regularization technique penalize whenever the output of the neural network doesn't match with any of these shapes. This type of special regularization has been used successfully for medical image analysis, where images are mostly corrupted because of the following reasons:

1. Low signal to the noise ratio.
2. Limited capabilities of the acquiring devices.
3. Motion artifacts.

Using the shape prior regularization in CNN networks, The predictions made are in accordance with that of the learned shape models of the undisclosed anatomy. As a consequence, these learned prior can help to improve the accuracy in the state-of-the-art neural network.

2.4 Summary

Regularization enforces some additional constraints while training deep learning models. These constraints aim to increase the prediction power and the generalization capabilities of the neural network. In this chapter, we discussed some basic regularization technique i.e. data augmentation, where we increase the amount of training data to make the model more generic and L1/L2 regularization, where we control the training parameters to prevent over-fitting.

In this report, the additional constraint took the form of prior knowledge about the shape of organ in human body. We already have this prior knowledge from the human anatomy. This knowledge could be used further, and added as an "regularization term" to improve the generalization of neural network models and so as the accuracy.

3 T-L regularization network

Girdhar et. al.[5] present the purported TL-embedding system, a mixture of a 3D auto-encoder to re-construct the voxel grid and CNN network to deduce the voxel grid from 2D images. The T and L here represent the architecture of neural network during training and testing phase. Their principle inspiration is to cater the following two problem:

1. To learn a generative representation in 3D.
2. To anticipate this representation from 2D images.

This network will be used to regularize our main approach (describe in chapter 4) for medical image segmentation and super resolution. In this chapter we'll discuss, how can we use T-L network for regularization of machine learning models and later on apply it for the medical image analysis tasks in ACNN networks.

3.1 Auto-Encoder

An auto-encoder (AE) [6] as described by Goodfellow et al. is a "neural network that targets to learn an intermediate description from which reconstruction of the original input is possible". The idea here is encoder part of neural network model is left uncompleted in lower dimension space, from this low dimension space we could further reconstruct the input again. This intermediate space is also know as latent space and aim here is to learn this space. This latent space could be represented in lower dimension, hence auto-encoder network are often used for dimesionality reduction.

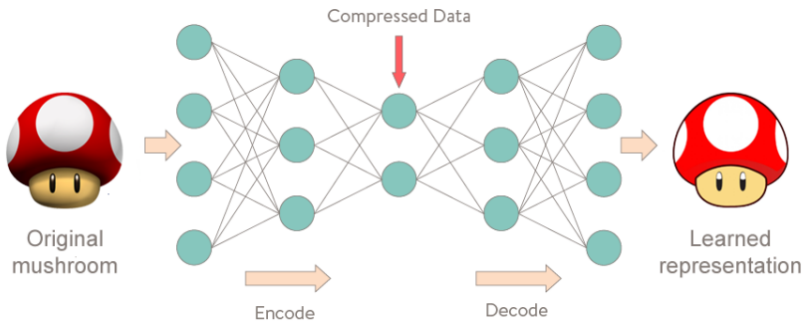

Figure 3.1: Auto-encoder network

This learning mechanism for auto-encoder enforces to select the most salient feature of the data. On the other hand, simply remove those which aren't useful or less useful.

The loss function for an auto-encoder network is explained as: $L_x(y_s, g(f(y_s)))$, where L_x is penalizing $g(f(y_s))$ being dissimilar from y_s.
g: Encoder components of the AE.
f: Decoder components of the AE.

3.2 Introduction to of T-L embedding Network

Task of T-L Network: To generate 3D representation of object from 2D images.
Basic assumptions are: Object representation must be

1. **Generative in 3D:** We should be able to reconstruct objects in 3D from it.

2. **Predictable from 2D:** We should be able to easily infer this representation from images.

3.3 T-L Network Architecture

Network architecture during Training T-L Embedding network [5] described by Rohit et al. have 2 main constituents, which are connected through 64-D vector embedding space:

1. An **auto-encoder network** maps a 3D voxel grid to the 64D embedding space, and decipher it back to 3D voxel grid. The input to auto-encoder network is 3D grid as shown in [**fig:3.2**] , which is the followed by four convolution layers. There is one fully connected layer after these convolution layers, which generated those embedded vectors. **Decoder** will then take this 64D embedded vector as input and re-map it to 3D voxels with 5 deconvolution layers.
   Parameters to train the auto-encoder network are described as below:
   - 3D convolution with stride 1.
   - Non linearity via parametrized ReLU.
   - Cross-Entropy as loss function.

2. A **discriminatively trained ConvNet** that learns the mapping from 2D image to the 64D embedding space. This mapping is learned by the lower part pf the T-Network as shown in [**fig:3.2**].

Network architecture during Testing At test time, The encoder part of the auto-encoder network is been removed and output of the image embedding network is connected directly to the decoder. The out of the decoder network is 3D voxel grid.

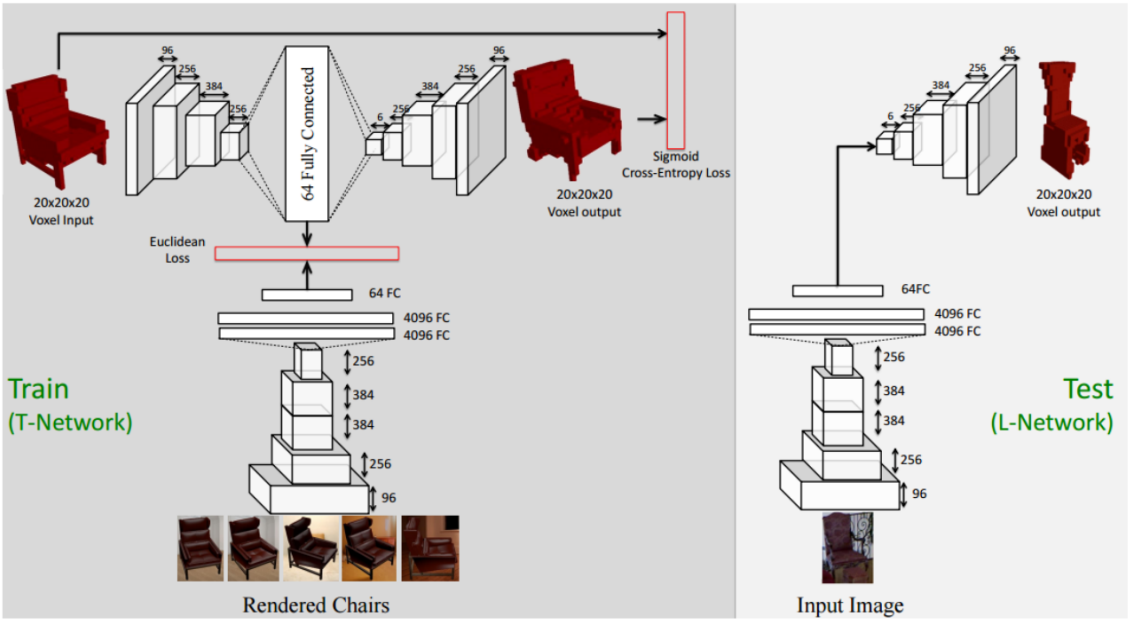

Figure 3.2: (a) T-network: At training time, the network takes two inputs: 2D RGB images and 3D voxel maps. (b) L-network: During testing, we remove the encoder part and only use the image as input.

3.4 TL Network Training

Training of T-L embedding network requires a lot of parameter tuning, therefore it is very difficult to train T-L network together from beginning. T-L network is trained in batches to simply the training process. As a result, we perform a **Three-stage Training** for this network.

1. **In the first stage**, the auto-encoder part of the network is been trained independent of Conv-Net. The network is trained end-to-end with the sigmoid cross-entropy loss and random initialization of the network.

2. **In the second stage**, the encoder part of the network produces the embedding of 3D voxel grid and convolution network is in-turn trained to regress those embedding. For the fast learning of the ConvNet, it is initialized with pre-trained weights from ImageNet.

3. **In the last stage**, the network is fine-tuned jointly with both the losses. ,

Most of the parameters are been updated in auto-encoder network because auto-encoder keep on updating its latent representation such that it would be pretty intuitive to regress that from 2D images.

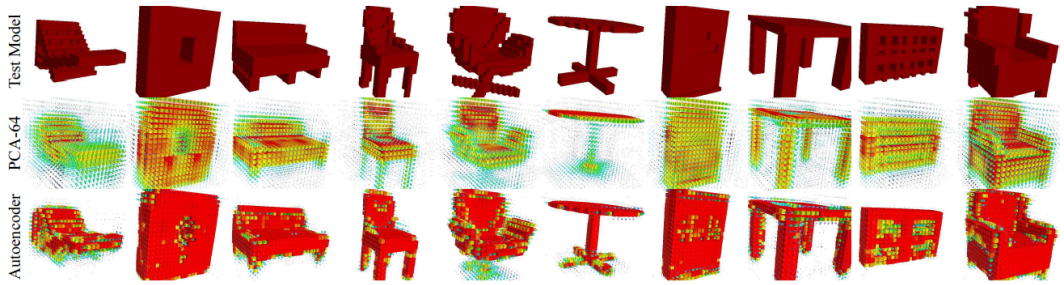

Figure 3.3: Predicted voxel are colored by level of prediction confidence i.e. from red to blue.

3.5 Conclusion

The qualitative results of T-L networks is shown in [**fig:3.3**]. From the result we can clearly see that auto-encoder networks are performing well over the other dimensionality reduction techniques like PCA. The auto-encoder networks are able to compute even the fine details such as the legs of the chairs, however PCA isn't performing good on those. The learned space with auto-encoder network is smooth i.e. computed reconstruction of linear interpolation accross two latent spaces. This further ensures that the representation is meaningful.

4 Anatomically Constrained Neural Networks

Oktay et. al. [7] described novel approach to incorporate the prior knowledge about the shape of organ in bio-medical images into the neural networks. Their work was mainly inspired by initial research in shape prior and image segmentation, which is based on PCA based active shape models and PCA based statistical models. Oktay et. al. used their framework for the two major task mentioned below for the bio-medical image processing:

1. Image Segmentation
2. Image Super- Resolution

4.1 Introduction

The constituents of "anatomically constrained neural networks" are image segmentation or super-resolution network and TL regularization network (as shown in [**fig:4.1**]). As the name suggest that, it extends the existing CNN network with extra regularization to make the predictions that are anatomically meaningful. The T-L network here is compose of stacked convolution auto-encoders [8], which learns the non-linear latent representation and the predictor network that regresses this learned latent space. The good part about the ACNN networks is that they are independent of the neural network architecture being used to regress that representation. Therefore, they could be merged with any modern deep learning approach to improve the overall accuracy without including any additional cost.

In ACNN models, regularization of neural networks using prior knowledge about the shape of the organs in bio-medical images is part of end-to-end learning which can be an extraordinary preferred standpoint. Also, using this regularization technique doesn't involve the usage of any post processing steps such as CRF (conditional Random Field)

4.2 Segmentation Network

The task of the segmentation is to take x = $\{ x_i \in \mathbb{R}, i \in S \}$ as input, where x = $\{ x_i \in \mathbb{R}, i \in S \}$ is the observed intensity image. The output of the segmentation network is $y_s = \{y_i\}_{i \in S}$, which represent an image of class labels with $y_i \in L = 1, 2,$ . . .C. In network based on CNN, we project this task as estimation of probability

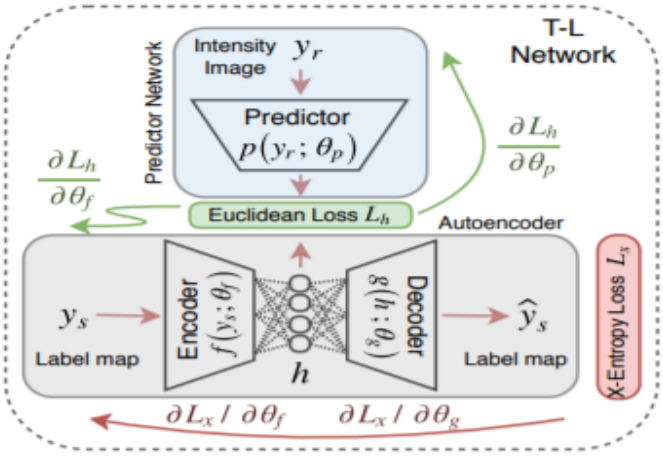

Figure 4.1: stacked convolution auto-encoders (AE) in gray combined with predictor network in blue to represent T-L network.

distribution function P(y_s |x).

The approximation of class densities P(y_s |x) involves the allocation of x_i to the probability of taking C classes. There are separate feature map for each class, resulting to total C sets of class feature maps fc. Each feature map is represented by learned non linear representation and then finally apply the softmax function to get the resulting class.

In the U-Net [9] as explained by Ronnenberger et. al. and DeepMedic models, we learn to map intensities to label i.e ϕ (x) : X $\longrightarrow$ L. We learns this mapping by optimizing the average cross-entropy loss of each class via stochastic gradient descent, where $L_x = \sum_{c=1}^{C} L_{(x,c)}$=1 depicts the cross-entropy loss of each class. This is shown in [**fig:4.2**].

Cross- Entropy Loss for segmentation:

$$L_x = -\sum_{c=1}^{C}\sum_{i\in S} \log(\frac{e^{f_{(c,i)}}}{\sum\limits_{j} e^{f_{(j,i)}}})$$

4.2.1 Segmentation with ACNN

Oktay et. al. [10] used AE as a regularization model and integrated that with the state of the art segmentation networks as in [11] to limit the prediction of class labels and pull the network towards more accurate output.

Using cross-entropy loss function for the optimization of neural networks fail to ensure the global consistency for segmentation network. This is due to the fact that cross-entropy loss function for the segmentation operate only on individual pixels i.e.

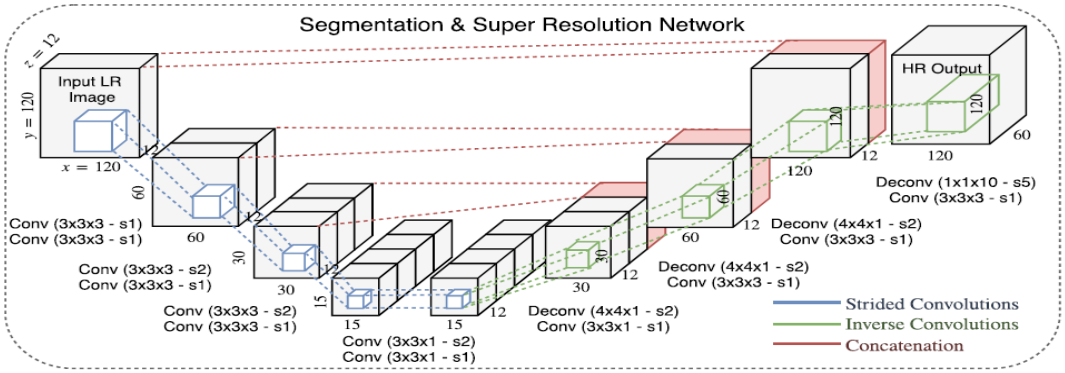

Figure 4.2: Neural network architecture for image segmentation (Seg) and super-resolution (SR)

doesn't consider the global context to account. This leads to the discrepancy in class label prediction.

To overcome this drawback, AE are used to suppress the dimension (e.g. 64 dimensions) of class prediction label maps. This low dimension representation contains information about segmentation and its underlying structure. ACNN constructs the objective function by comparing the AE-based latent space of ground-truth and network prediction, as shown in [**fig:4.3**]. ACNN-Seg training objective function is described as below:

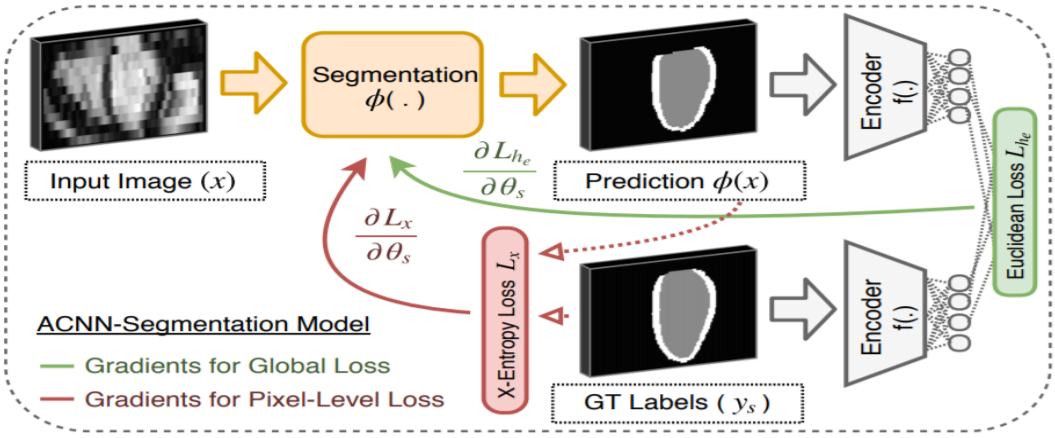

Figure 4.3: Training methodology of proposed anatomically constrained convolution neural network (ACNN) for image segmentation tasks.

Segmentation Loss with shape regularization:

$$Loss = min\theta_s(L_x(\phi(x;\theta_s),y) + \lambda_1.L_{h_e} + \frac{\lambda_2}{2}||w||_2^2)$$

L_x: Cross-entropy
θ_s: All the trainable parameters of segmentation model
λ_1 , λ_2: Weight of 'shape regularization loss' and 'weight decay term' L_{h_e}: Penalize, if the latent space representation of predicted label is different than that of ground-truth.
$\frac{\lambda_2}{2}||w||_2^2$: Limits the number of free parameters while training model to avoid over-fitting.

4.3 Super-Resolution Network

As per the super resolution network as explained by Oktay et. al.[10], mapper function is learned $\phi : X \longrightarrow Y$ which approximate a high-resolution image $\hat{y_r} = \phi(x : \theta_r)$. where, θ_r represents model parameters such as convolution kernels and batch normalization statistics. These parameters are tuned by minimizing the smooth l_1 loss across the ground-truth HR image and the respective prediction. The smooth l_1 norm is determined as below:

Ψ_{l_1}: Smooth l_1 loss defined as { $0.5k^2$ if |k|<1 , |k| - 0.5 otherwise }
SR Training Objective is:

$$\sum_{i \in S} \Psi_{l_1}(\phi(x_i; \theta_r) - y_i)$$

Super-resolution is generally an ill-posed problem, because there could be many solutions to a single problem. However, all of them wouldn't be correct solution to the problem. ACNN is used to limit the possible solutions via regularizing the correctness of result.

4.3.1 SR Framework with ACNN

The output of the super-resolution network is intensity image , therefore shape encoder AE will not work here. To avoid this problem, we extends the standard denoising encoder to T-L regularization model and to achieve that AE is integrated with predictor network as show in [**fig:4.4**]. This predictor network do the mapping of input image to the low dimensional non-parametric representation of anatomy, represented symbolically as p(x) : X $\longrightarrow$ H. This mapping is learned by AE networks.

To regularize the SR model, hidden space is been learned from both image label space Y and image intensity space X. The finally penalize on the basis of euclidean loss between the two mappings.

Super Resolution Loss using shape regularization:

$$Loss = min\theta_s(\Psi_{l_1}(\phi(x; \theta_r) - y_r) + \lambda_1 . L_{h_p} + \frac{\lambda_2}{2}||w||_2^2)$$

Ψ_{l_1}: Smooth l_1 loss defined as { $0.5k^2$ if |k|<1 , |k| - 0.5 otherwise }
θ_r : All the trainable parameters in SR model.

λ_1 , λ_2: Weight of 'shape regularization loss' and 'weight decay term' L_{h_p}: Penalize, if the latent space of the intensity image doesn't map with that of the ground-truth labels.
$\frac{\lambda_2}{2}||w||_2^2$: Limits the amount of features required in deep model to avoid over-fitting.

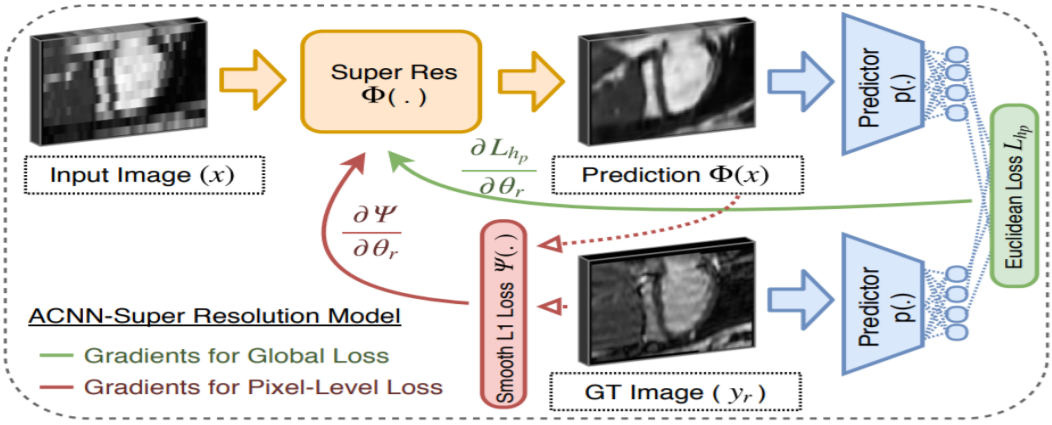

Figure 4.4: Training methodology of the proposed anatomically constrained convolution neural network (ACNN) for image Super Resolution tasks.

4.4 ACNN: Advantages

ACNN network is been widely used for image analysis because of the two major reasons as mentioned below:

1. The regularization network mentioned here can be used for **any other image analysis task** other than just image segmentation or image SR. This global information about the anatomy is learned at the training time. The regressor network of image segmentation or image SR is then provided with this extra knowledge. As a result, the output is meaningful and more robust against the image artifacts.

2. ACNN use T-L embedding networks for regularization, which **generalizes** pretty well. In other word, we could say that the learned representation could be predicted easily for both the intensity space and the label space because of the joint training of AE and predictor.

4.5 Results

The results are been shown in [**fig:4.5**] and [**fig:4.6**]. The results depicts the usefulness of introducing shape priors in segmentation models. We can see that using ACNN

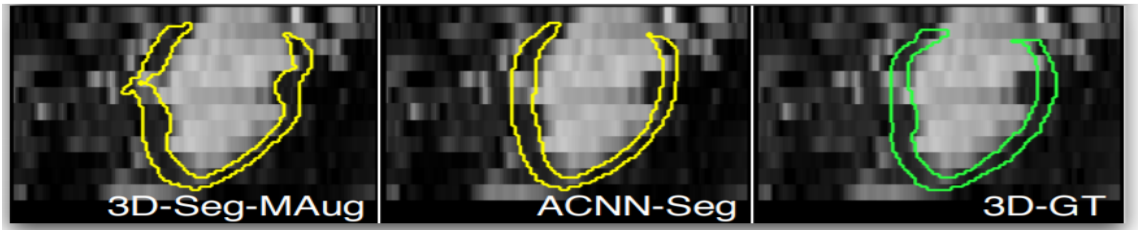

Figure 4.5: Segmentation results on two different 2D stack cardiac MR images.

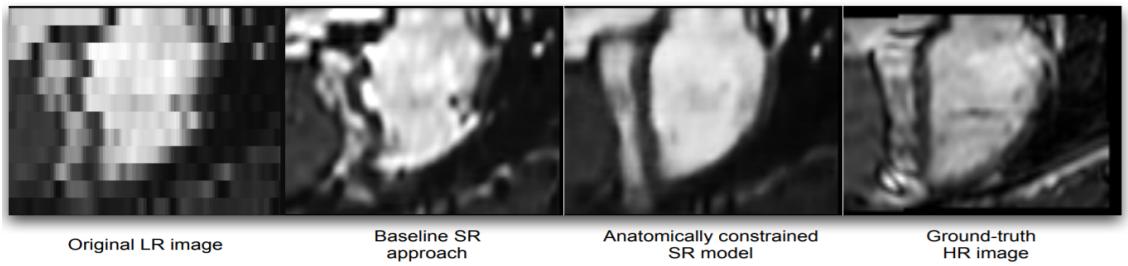

Figure 4.6: Image super-resolution (SR) results.

networks we could tackle false-positive detection and motion-artifacts. Also, It can be seen from the [**fig:4.5**], that ACNN is insensitive to slice misalignment as it is anatomically constrained and it makes less errors in basal and apical slices compared to the 2D-FCN approach.

4.6 Conclusion

Training objective has been altered by the regularization term. This is been observed that by tuning new objective function, neural network are making prediction that are confirming the leaned shape model of the hidden anatomy (also referred as image priors).

ACNN are particularly useful, when images acquired are of worse quality and requires a lot of pre-processing. Inspite of that, the results obtained using such images are quite useful and are not totally senseless.

Another interesting fact is that in-spite of really complicated architecture, ACNN is computationally efficient in terms of run-time as compared to the SR-CNN model by a factor of 5. The reason to this is ACNN-SR do feature extraction in low dimension latent space while the other one do it on original representation.

5 GridNet

The GridNet [12] as described by Zotti et. al. accompanies with installed shape priors. The objective function is specific only to the cardiac anatomy. This technique process raw MR pictures with no manual preprocessing or potentially image trimming. The used convolution neural network (CNN) learns both **high-level** features (to distinguish heart from any other organ in human body of same shape or size) and **low-level** features (to extract the accurate segmentation results). Both types of features are learned with "Grid" architecture, which is just an extension of existing segmentation network for medical images i.e. U-net [9]

5.1 Introduction

The method described in this chapter using gridnet is the first CNN method that is designed to segment different components of the human heart i.e. **left ventricular/LV, right ventricular/RV and myocardium/MYO** without any external segmentation method. The incorporation of the prior knowledge about the shape is in-build in this network.

Some of the pre-existing techniques for cardiac segmentation are described in [13]. A few of them uses conventional techniques like Hough transform and few of them are based on most recent deep learning methods and CNN.

5.2 GridNet and shape prior

5.2.1 Input and processing in GridNet

The input to the GridNet is a 3D L X B X H raw MR image X. The aim of the GridNet is to segment the RV, the MYO and the LV of 3D input image.

To achieve this, 3D label map T also of size L X B X H is predicted. The voxels v = (i; j; k) of predicted label map comprises of label $T_v \in$ (LV, RV, MYO and Back), The “Back” here stands for classes other than LV, RV and MYO.

As described in ACDC (Automatic Cardiac Diagnosis Challenge) structure, 3D input image i.e. X is sequence of axis aligned slices commences from the mitral valve and ends at the apex. The role of the shape prior here is same as described in chapter 4 i.e. it enforces clinically plausible result. Shape prior in GridNet contains informa-

tion about the corresponding position of the LV, RV and MYO.

The prime task here is to adjust shape prior over the input info X precisely. To do this, we enlist S on X by deciphering the focal point of S on the cardiovascular focal point of mass (CoM) c of X. To decide c, there is independent relapse module in Gridnet.

5.2.2 Specific shape priors

The shape prior S in GridNet encodes the probability of belonging 3D voxel v = (i; j; k) to a certain class (Back, LV, RV, or MYO). To estimate this probability, pixel-wise empirical ratio for every class is been computed on ground-truth label fields T_i of the training dataset:

$$P(C|v) = \frac{}{N_t} \sum_{l=1}^{N_t} 1_c(T_{i,v})$$

$1_c T_{i,v}$: Indicator function return 1 when $T_{i,v}$ = C and zero otherwise
N_t: the total number of training images.

The information obtained above for the shape prior is encoded into a 3X20X100X100 volume S, where 3 here represents the 3 classes (RV, MYO, LV), 20 represents the number of interpolated slices and 100 X 100 is the inplane size.

5.2.3 Objective functions for GridNet

$L_T = -\gamma_T \sum_{l=1}^{4} \sum_{v} T_{i,l,v} \ln \hat{T}_{l,v}$ (Cross-entropy of predicted labels)
$L_C = -\gamma_C \sum_{l=1}^{4} \sum_{v} C_{i,l,v} \ln \hat{C}_{l,v}$ (Cross-entropy of predicted contour)
$L_c = \gamma_c ||c_{i,w} - \hat{c}_i||^2$ (Euclidean distance between predicted CoM $c_{i,w}$ and true CoM)
$L_w = \gamma_w ||w||^2$ (Prior loss)

Total Loss or Objective function:

$$L = \sum_{i} (-\gamma_T \sum_{l=1}^{4} \sum_{v} T_{i,l,v} \ln \hat{T}_{i,l,v} - \gamma_C \sum_{l=1}^{4} \sum_{v} C_{i,l,v} \ln \hat{C}_{l,v} + \gamma_c ||c_{i,w} - \hat{c}_i||^2) + \gamma_w ||w||^2$$

i: Corresponds to slice.
l: Classes (4 classes: RV, LV, MYO, Back)
v: Pixel location
$T_{i,l,v}$: True probability that for slice i pixel v is in class l
$\hat{T}_{i,l,v}$: Output of our model for slice i pixel v is in class l
$C_i, \hat{C}_i$: Contour extracted from $T_i, \hat{T}_i$
γ_T, γ_C: Constants

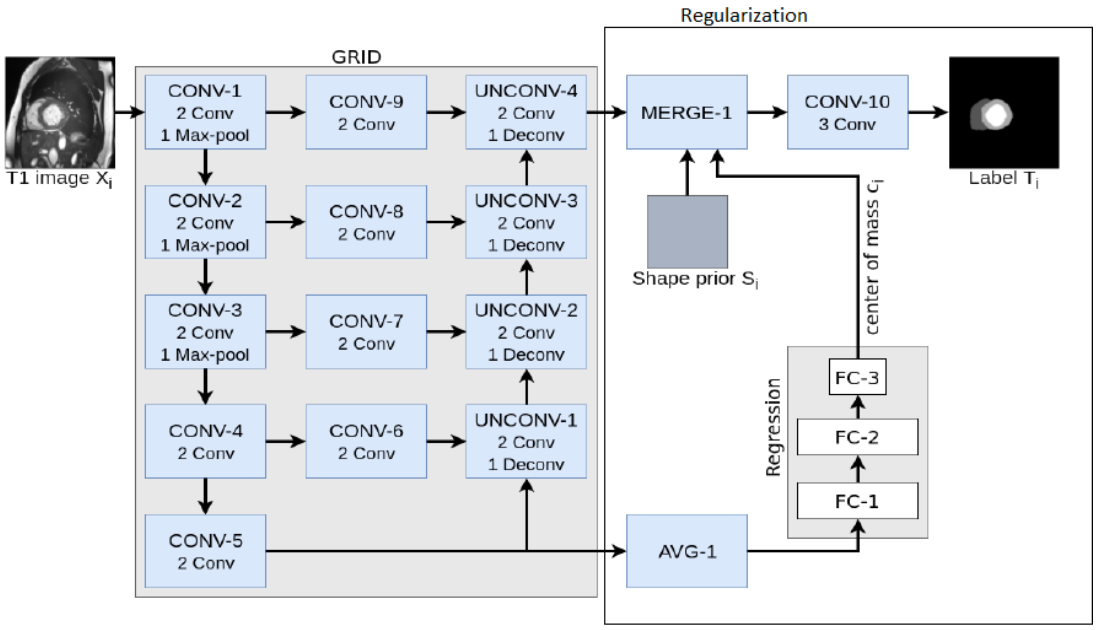

Figure 5.1: Network architecture of GridNet

5.3 Network Architecture

The network architecture of GridNet consisting of 3 columns and 5 rows and looks similar to that of Grid as shown in [**fig:5.1**]. Model's input are 256^2 cardiac image acquired via MRI X_i and the shape prior S_i of same slice, while the output is the center of mass c_i and a label field T_i.

The common issue with the acquisition of MRI cardiac images is due to its slow acquisition and respiratory motion of the heart the location of slice got shifted between the consecutive slices. Therefore, it is quite difficult to handle the entire 3D volume. To deal with the situation, the network is inputted with 2D slices and reshape the 3D volume T_i by piling up 2D label fields.

Global context in terms of features can be extracted as we deep down into the network. As high level features of image available at CONV-5 layer, image's X_i cardiac CoM, ci can be predicted by using that layer.

There are four different convolution layers without max-pool in second column of the network. All of these computes features at different resolution. As further column in the network are traversed (i.e. the last column), features at various resolution are been accumulated. Therefore, at the end UNCONV-4 contains features from both global and the local level details. This in turn are use to segment the MRI cardiac image.

We could also compare this grid structure with that of U-Net [9] except the below two differences

1. GridNet contains additional middle layer comprise of CONV-6 to 9.
2. CONV-5 features are been used to compute the CoM c_i.

To summarize, the network consisting of two major modules namely:
- **Center of mass c_i estimation**: This done with regression module after CONV-5 layer.
- **Label field T_i estimation**: To achieve this, MERGE-1 is used to register the output of UNCONV-4 on shape prior S using CoM c_i obtained above. At the end, CONV-10 converts the feature maps obtained from MERGE-1 to single 4D output.

5.4 Conclusion

The GridNet is the CNN method, which is more generic form of U-Net [9] architecture for medical image segmentation. Shape priors are been imposed onto the intensity image as part of the network itself. It appends the shape prior information with the help of regression method, unlike PCA based used in ACNN.

The only limitation to this architecture is that, they could only be used for MRI cardiac segmentation. In addition, this is been verified by the experiments that this approach is well suited for fully automatic clinical tool.

6 Conclusion and Discussion

6.1 Recap

Prior knowledge about the anatomy provided guidance and robustness to models. This is acting as an additional regularization term, which helps to generalized out model better. The neural network models are performing better after incorporating this regularization term.

Two major techniques are discussed in this report to incorporate this knowledge in NNs models:

- GridNet: Registration of shape prior on top of the input image using regression based method.
- Anatomically constrained Neural Networks: This is based on PCA based latent representation of shape prior and compare this representation with input.

As already discussed in chapter 4, that the regularization using shape prior is independent of underlying network architecture. Therefore, we could extent this technique to other image analysis tasks where prior knowledge could be useful to improve results.

6.2 Conclusion and discussion

In this report, we have discussed two different neural network architecture that incorporate these special regularization into their objective function. One is based on latent space or PCA based representation using auto-encoder network and the other one is based on registration of shape prior on intensity image using regression module.

The advantage of ACNN based networks are that they are independent of underlying neural network i.e. we could impose the regularization with any latest deep model and improve the results. However, the number of parameters required for this approach are very high and so is the computational complexity. Additionally, we could use this approach only for the generic representation of the ground truth.

On the other hand, regularization of GridNet using the registration of shape prior is quite fast and computationally efficient. However, we could use this technique only for the segmentation of MRI cardiac images.

In-spite of all the cons and pros of using shape prior regularization in image segmentation and image networks. There's huge improvement in accuracy of results. The new objective functions ensure the meaningful representation of different body organs, which further improves the possibility of detecting any foreign cell or disparity with human organs. As a result, we could better detect any disease or unwanted growth at very early stage, even if image acquisition quality is poor.

List of Figures

Bibliography

[1] Hao Chen et al. "DCAN: Deep contour-aware networks for object instance segmentation from histology images." In: *Medical Image Analysis* 36 (2017), pp. 135–146.

[2] Tobias Heimann and Hans-Peter Meinzer. "Statistical shape models for 3D medical image segmentation: A review." In: *Medical Image Analysis* 13.4 (2009), pp. 543–563. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2009.05.004. URL: http://www.sciencedirect.com/science/article/pii/S1361841509000425.

[3] Wenjia Bai et al. "A Probabilistic Patch-Based Label Fusion Model for Multi-Atlas Segmentation With Registration Refinement: Application to Cardiac MR Images." In: *IEEE Transactions on Medical Imaging* 32 (2013), pp. 1302–1315.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[5] Rohit et al. Girdhar. "Learning a Predictable and Generative Vector Representation for Objects." In: *European Conference on Computer Vision. Springer, pp. 484–499.* 5 (2016).

[6] Guillaume Alain and Yoshua Bengio. "What regularized auto-encoders learn from the data-generating distribution." In: *The Journal of Machine Learning Research 15.1, pp. 3563–3593.* (2014).

[7] O et al. Oktay. "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation." In: *IEEE transactions on medical imaging, ISSN: 1558-254X, DOI:10.1109/TMI.2017.2743464* 1 (2018).

[8] Jonathan Masci et al. "Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction." In: *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I.* ICANN'11. Espoo, Finland: Springer-Verlag, 2011, pp. 52–59. ISBN: 978-3-642-21734-0. URL: http://dl.acm.org/citation.cfm?id=2029556.2029563.

[9] O. Ronneberger, P.Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. LNCS. (available on arXiv:1505.04597 [cs.CV]). Springer, 2015, pp. 234–241. URL: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.

[10] O. Oktay et al. "Multi-input cardiac image super-resolution using convolutional neural networks." In: *MICCAI* 7 (2016), pp. 246–254.

[11] E. Shelhamer J. Long and T. Darrell. “Fully convolutional networks for semantic segmentation.” In: *Proc. CVPR, 2015, pp. 3431–3440.* 4 (2015).

[12] Zotti Clement et al. “"GridNet with automatic shape prior registration for automatic MRI cardiac segmentation".” In: 4 (2017). URL: `https://arxiv.org/pdf/1705.08943.pdf`.

[13] C. Petitjean et al. “Right ventricle segmentation from cardiac MRI: A collation study.” In: *Med. Image Anal., vol. 19* (2015), p. 187.