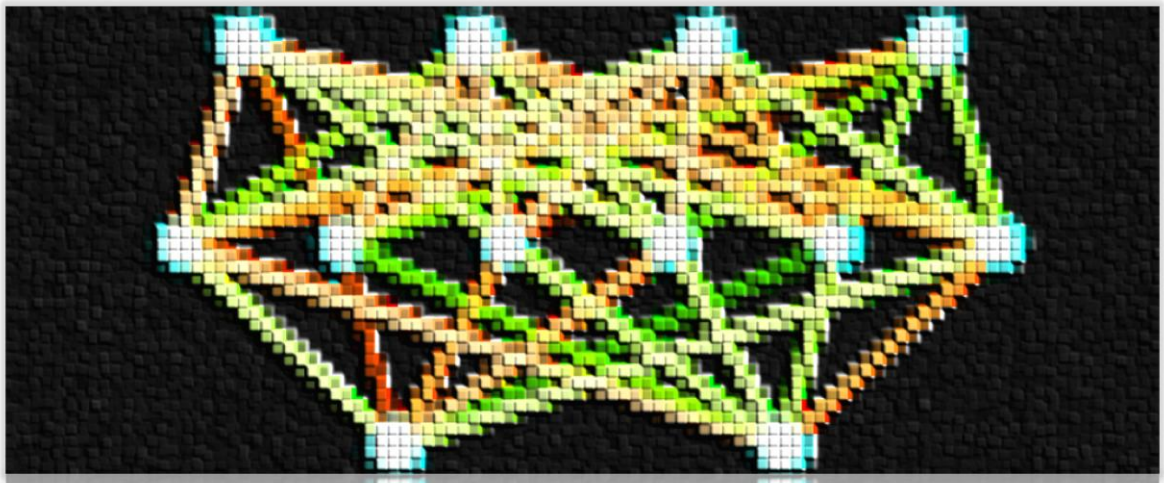

EXTRACTING HIGH-DIMENSIONAL FEATURES FROM MEDICAL IMAGES BY UTILIZING DEEP LEARNING TECHNIQUES

MASTER OF SCIENCE

DEPARTMENT OF INFORMATICS ENGINEERING

SCHOOL OF ENGINEERING

TECHNOLOGICAL EDUCATIONAL INSTITUTE OF CRETE

July 2018

Heraklion

by Eleftherios Trivizakis

SUPERVISOR | KONSTANTINOS MARIAS, PHD

Abstract

Machine Learning, as a field of computer science started in the early 1960s. Recently, with the rise of high-throughput computing and the massively available data, advanced machine learning techniques have led to unprecedented results in image processing and computer vision analysis applications. This was due to the introduction of novel architectures such as recurrent, deep belief, convolutional networks and other deep architectures, as well as the availability of more sophisticated hardware or cluster computing, better training methods and strategies, efficient regularization and normalization algorithms.

A large number of image processing and computer vision applications have seen significant benefits from the application of advanced machine learning techniques including image classification, text-to-image retrieval, object recognition, enhancement, registration, segmentation and generation. Regarding, Medical Imaging deep learning applications such as automated organ or object segmentation, lesion classification or detection and image quality enhancement have achieved significantly advanced results, performing near or even better than a human expert.

In this master thesis a 2D and a 3D CNN has been developed for extracting multi-dimensional medical image features for discriminating between primary vs metastatic malignant liver lesions. The proposed network consists of four consecutive strided 3D Convolutional layers with 3x3x3 kernel size and ReLU as activation function followed by a fully-connected layer with 2048 neurons and a Softmax layer for binary classification.

A dataset comprised of 107 DW-MRI scans was used for training and validation in both networks. In particular, 2D and 3D CNNs were designed for liver cancer differentiation and trained by the same dataset. The results demonstrated a superior classification performance of the 3D over the 2D with 93.9% vs 65% and 95% vs 67% tissue classification accuracy and sensitivity respectively.

These results suggest that there are potential benefits of Deep Learning on diffusion MRI in enhancing the diagnostic information towards more precise, personalized health care of liver cancer patients. Early image-based information regarding primary-metastatic pathology can optimize therapy decisions early and spare patients of unnecessary therapy.

Περίληψη

Η Μηχανική Μάθηση ως κλάδος της Επιστήμης Υπολογιστών ξεκίνησε την δεκαετία του 1960. Πρόσφατα οι εξελιγμένες τεχνικές μηχανικής μάθησης, υποβοηθούμενες από την άνοδο της υπολογιστικής ισχύς και την διαθεσιμότητα μεγάλου όγκου δεδομένων, έχουν οδηγήσει σε εξαιρετικά αποτελέσματα για εφαρμογές επεξεργασίας εικόνας και υπολογιστικής όρασης. Αυτό οφείλεται στην εισαγωγή νέων αρχιτεκτονικών όπως τα συνελικτικά νευρωνικά, στοιβαγμένα πολύ-επίπεδα δίκτυα και άλλες βαθιές αρχιτεκτονικές δικτύων, καθώς και στην διαθεσιμότητα πιο εξελιγμένου υλικού, τις βελτιωμένες τεχνικές εκπαίδευσης των δικτύων και ομαλοποίησης/κανονικοποίησης πληροφορίας.

Ένας μεγάλος αριθμός εφαρμογών επεξεργασίας εικόνας και υπολογιστικής όρασης έχουν ωφεληθεί από αυτές τις τεχνικές μηχανικής μάθησης περιλαμβάνοντας την κατηγοριοποίηση εικόνας, αναγνώριση αντικειμένου, βελτιστοποίηση, καταχώριση-αποτύπωση και κατάτμηση εικόνας. Σχετικά με τις εφαρμογές Ιατρικής Απεικόνισης όπως η αυτοματοποιημένη κατάτμηση οργάνου ή αντικειμένου, κατηγοριοποίηση ή ανίχνευση οργανικής βλάβη, αυτές έχουν σαφώς εξελιχθεί αποδίδοντας κοντά ή και καλύτερα από τον αντίστοιχο ειδικό.

Στην παρούσα μεταπτυχιακή διατριβή, έχουν υλοποιηθεί δυο Συνελικτικά Δίκτυα (ένα 2-Δ και ένα 3-Δ) για την εξαγωγή πολυδιάστατων χαρακτηριστικών από ιατρικές εικόνες, ώστε να είναι δυνατή η κατηγοριοποίηση μεταξύ πρωτογενών ή μεταστατικών κακοηθών όγκων. Το προτεινόμενο δίκτυο αποτελείται από τέσσερα διαδοχικά 3-Διαστάσεων συνελικτικά επίπεδα με μέγεθος πυρήνα 3x3x3 και συνάρτηση ενεργοποίησης ReLU ακολουθούμενη από ένα πλήρως συνδεδεμένο επίπεδο με 2048 νευρώνες και κατηγοριοποιητή Softmax.

Για την εκπαίδευση και αξιολόγηση των δικτύων τα δεδομένα που χρησιμοποιήθηκαν και για τα δύο δίκτυα περιλαμβάνουν 107 DW-MRI σαρώσεις. Τα 3-Δ δίκτυα παρουσιάζουν ανώτερες επιδόσεις σε κατηγοριοποίηση ιστού έναντι των 2-Δ δικτύων με 93,9% προς 65% και 95% προς 67% για ακρίβεια και ευαισθησία κατηγοριοποίησης αντίστοιχα.

Τα αποτελέσματα αυτά συνηγορούν στο ότι υπάρχουν σημαντικά δυνητικά οφέλη με την ενίσχυση της διαγνωστικής πληροφορίας προς ακριβέστερη, εξατομικευμένη περίθαλψη σε ασθενείς που πάσχουν από καρκίνο στο συκώτι. Η έγκαιρη, βασισμένη σε ιατρικές εικόνες, πληροφόρηση σχετικά με την πρωτογενή ή μεταστατική παθολογία του ασθενή μπορεί να βελτιστοποιήσει τις θεραπευτικές αποφάσεις νωρίς ώστε να αποφευχθούν άσκοπες θεραπείες.

Table of Contents

List of Figures

List of Tables

List of Equations

Acknowledgements

Firstly, I would like to thank prof. dr. Konstantinos Marias for his supervision throughout the research that led to this thesis at the Technological Educational Institute of Crete. At Foundation of Research and Technology Hellas I would also like to thank George Manikis and Katerina Nikiforaki for their valuable contribution, feedback and expertise related to the clinical part of the research.

July 2018

Eleftherios Trivizakis

1. Introduction

Many researchers are predicting that Artificial Intelligence (AI) will eventually predominate human intelligence in the near future. In fact, specialized AI already outperforms even human experts in some tasks. Throughout the literature a new highly expanding topic is on the rise. fast. Deep Learning (DL) on image analysis minimized the error from 25% (ImageNet Challenge 2011) to less than 5% in 2017. This unprecedented performance led to fast adaptation on domains other than "natural" image applications. In particular, recent DL studies on medical imaging has demonstrated promising results for different tasks and diagnosis objectives.

Liver cancer tissue differentiation is a challenging task due to the high variability in shape and dispersity of malignant tissue across the organ. Clinicians cannot make a conclusive diagnosis only from visual clues without an invasive procedure. Different types of cancer determine the treatment path that will be followed. In this thesis we explore the potential of implementing a deep Convolutional Neural Network to predict the tissue cancer type: primary or metastatic resulting in increased survivability for the patient.

1-1. Research Questions

In this study we apply deep learning techniques and models on medical image analysis and evaluate the quality of predictions for liver cancer tissue discrimination between primary and metastatic. The following questions are studied:

- Can Deep Learning assist in a challenging diagnosis problem such as tissue differentiation?
- Can data augmentation approaches contribute to better models despite the lack of data for medical image analysis?
- How Deep Features extracted from different Convolutional Neural Network architectures compare?
- Which architecture is more suitable or performs better for liver cancer tissue discrimination?

We propose a new methodology and baseline for liver cancer tissue discrimination based on strided 3D Convolutional Neural Networks and data augmentation techniques for alleviating the class imbalance in the dataset.

1-2. Thesis Outline

This thesis is comprised of six chapters:

1. Introduction
2. Deep Learning in Computer Vision
3. Medical Image Analysis
4. Convolutional Neural Networks
5. Proposed Research
6. Conclusion

In Chapter 2 we provide an in depth and comprehensive perspective of publications for deep learning image analysis presenting the state-of-the-art architectures, transfer learning approaches, unsupervised trained models, some insights and criticisms related to this new topic of Artificial Intelligence. The 3rd Chapter is the literature review to medical image analysis mostly focused in deep learning. In Chapter 4 we present some essential mechanisms and concepts related to Convolutional Neural Networks. An extensive analysis of the different architectures and approaches, proposed methodology and quantitative evaluation of the models are provided in the 5th Chapter. Finally, in Chapter 6 we discuss the research findings and the contribution of deep learning in medical image analysis, particularly in liver cancer tissue discrimination.

2. Deep Learning in Computer Vision

2-1. Learning Algorithms

Deep learning, as a topic of artificial intelligence and machine learning, enables adaptation to a domain's needs without providing explicit details about the problem itself or the solution.

There is a wide variety of deep architectures distinguished into unsupervised and supervised learning models including autoencoders [1] [2] [3], convolutional deep belief networks [4] [5], artificial neural networks [6], recurrent networks [7] [8] [9] [10], convolutional networks [11] each one specialized in a specific domain or a type of dataset.

The Convolutional Neural Networks (CNN) [11] are the current state-of-the-art for image related tasks like classification [12] [13] [14] [15], detection [16] [17], segmentation [18] [19] [20] [21] and even image generation [22]. A large number of labeled images for training is required to achieve such an outstanding performance. For instance, a typical benchmark dataset like CIFAR-10 [23] consists of tens of thousands natural color images, 10 classes with 5000 samples per class exclusively for training.

This infamous black box procedure does not consist of propagation of random patterns throughout the network as it is a common belief. To the contrary visualizing the feature maps [24] reveals the causality of the strong discrimination power of their representation with properties like compositionality and invariance as layers progress unveiling the local dependences, an important feature for classification.

2-2. Classic Image Processing Algorithms

In recent years researches advanced the "shallow" or the classic image processing algorithms [25] for image classification by introducing the main Fisher Vector pipeline with processes like subsampling, spatial information encoding [26] [27], dense SIFT [28], PCA [29], normalization [30], Fisher encoding with Gaussian Mixture Model (GMM) [31] and linear classifiers. Fisher Vectors [32] incorporate statistical information about the distribution of descriptor combined with the gradient vector from the GMM, retaining a compact feature vector for computational performance without sacrificing representative expression.

On the contrary, in [33] the authors present a Deep Fisher Kernel layer approach where each process of the pipeline is interconnected to solve the combined convex shaped problem of GMM convergence and SVM classification as an end-to-end feedforward layer. This approach deviates from the typical implementations of Fisher pipeline in literature and draw inspiration from the end-to-end nature of learning algorithms.

2-3. Supervised - Deep Learning

The Deep Learning (DL) architectures are far superior [34] [35] in computer vision tasks at the same dataset compared to hand-crafted algorithms in terms of performance, pipeline efficiency, compactness of representations and marginally faster inference.

Depth is beneficial [36] for learning models in terms of performance, yielding lower test errors, a sign of good generalization, compared to "shallow" algorithms. Deep and "thin" models [37] with more layers and less learned kernels can learn generalized representations, transferable to other domains or specialize recognition at subclass, instance retrieval from the source domain.

2-3-1. Classification Networks

The most influential architecture in DL is no doubt AlexNet [38] with its 8 layers, 1000 classes and its performance, top-5 error 15.4% vs the previous 26.2% in 2012 ImageNet Large-Scale Visual Recognition Challenge, managed to propel the new era of computer vision.

GoogLeNet [39], a network built on Inception blocks, is focused on reducing the number of parameters, boosting generalization, introducing a sparse layer architecture instead of multiple wide fully-connected layers. Even though less fully-connected units mean even less computational operations, today's hardware and software is optimized mostly for dense matrix

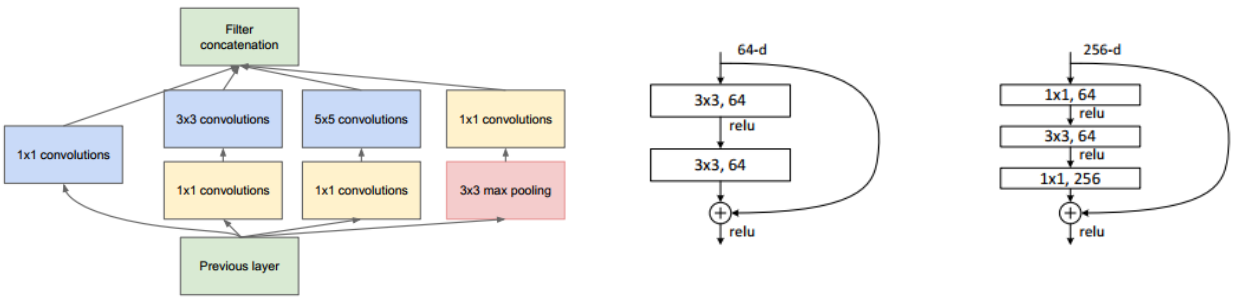

Figure 1 – Inception [39] and Residual Blocks [41]

multiplication limiting the potential gains of sparsity.

Routing information across very deep networks [40] is important and crucial for their successful training. Deep Residual Networks [41] [42] and Highway Networks [43] can reach extreme depth addressing the degradation problem in a similar way. This is achieved by providing shortcut connection between blocks in favor of introducing additional layers, leading to gains in terms of representation power. Other benefits of residual learning include easy optimization, compact models with fewer parameters at each layer and lower complexity. Also, migrating architecture characteristics of residual networks to more traditional approaches offers gains related to the training process and generalization as suggested in Inception v4 [44] network.

2-3-2. Segmentation Networks

Semantic segmentation, the process of performing per pixel classification separating the distinct objects in an image, is a fundamental problem in computer vision. Simple patch-based approaches like in [45] utilizes pre-trained classification networks, remove the layers related to that task, re-purpose the convolutional part as a feature extractor and append new additional layers for upscaling the predicted segmentation mask or calculate the probability scores.

Powered by a pre-trained VGG-16 model in [19] [46], the authors implement a Fully Convolutional Network (FCN) with no fully-connected layers. Instead up-sampling, deconvolutional or unpooling, layers output a corresponding to the input size, segmentation mask. Fine-tuning the new networks is required for better adaptation to a custom dataset providing more accurate predictions during inference.

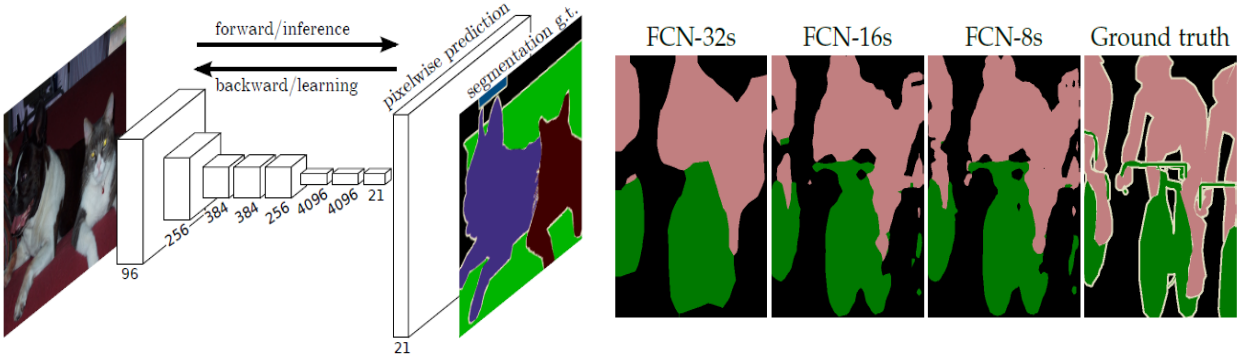

Figure 2 - FCN Pixel-wise Classification [19]

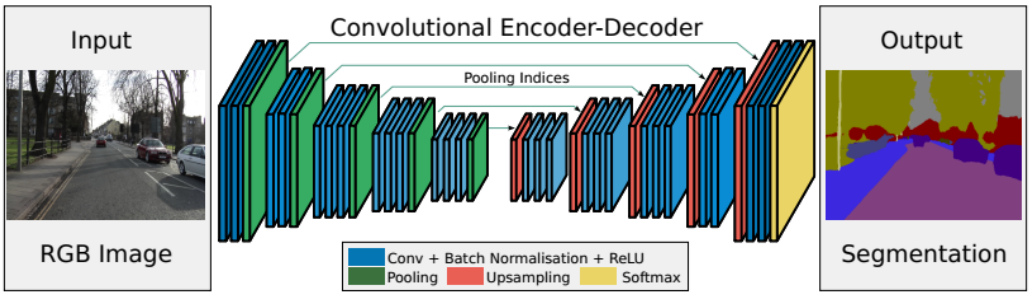

Figure 5 - SegNet Architecture [20]

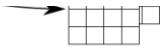

Figure 3 – Illustration of FCN architecture [19]

The refinement of these predictions is also subjected to the implementation of the skip connection architecture. Three types of those connections are suggested: FCN8, FCN16 and FCN32 (fine to coarse) depending on the computational cost and how much information from shallower (finer) layers is combined with the standard (coarser) output of the up-sample component of FCN.

Instance Sensitive FCN (IS-FCN) [47] extends the simple FCN by discriminating among instances of the same class utilizing the local coherence property and exploit the differences among predictions. A modified VGG-16 is used as feature extractor with no fully-connected layers and reduced stride. The instance-sensitive segmentation mask is the combination of two fully convolutional branches that produce multiple segmented instances describing the relative position, with convolutional layer setup 1x1 and 3x3, and their score, with 3x3 and 1x1 convolutions.

The architecture of SegNet [20] resembles the fundamentals of FCN but with one key difference, as a decoder is utilized only the max-pooling indices from the encoder and not the

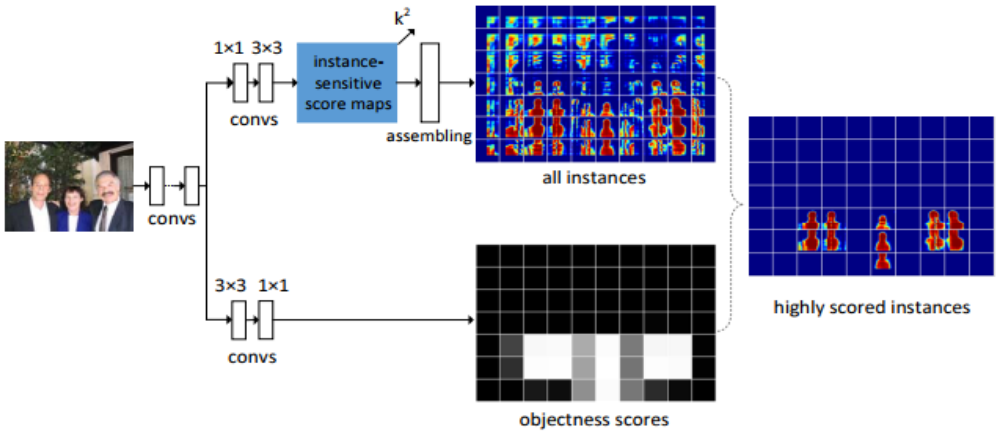

Figure 4 - Instance-Sensitive FCN [47]

whole feature maps leading to lower resource consumption and learnt parameters. This tradeoff though has an impact on the quality of the predictions compared to FCN performance, making the SegNet a compact, less demanding and less performing alternative to FCNs or U-Net.

DeepLab [48] architecture introduces dilated convolutions for the up-sampling part of the network offering larger fields of view with smaller kernels. In addition, fully-connected Conditional Random Fields (CRFs) alleviate the lack of connectivity among neighbor pixels produced by the deep network resulting in refined segmentation masks.

Most of the deep segmentation architectures are implemented as pixel-wise classifications with dependence on human expert knowledge by requiring accurate per slice delineations of regions of interest. This makes the data-related processes time and cost consuming. The authors of [49] propose a deep network for semantic segmentation using semi-supervised learning with weak annotations and few fully labeled data.

2-3-3. Multi-Stream Networks

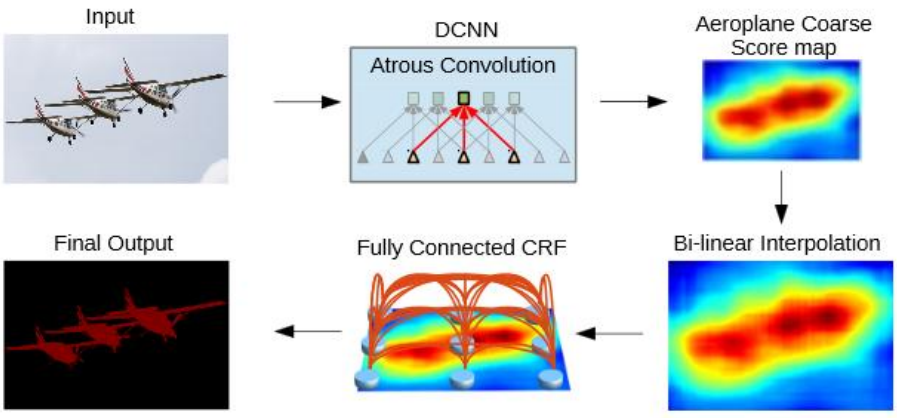

Figure 6 - DeepLab Pipeline [48]

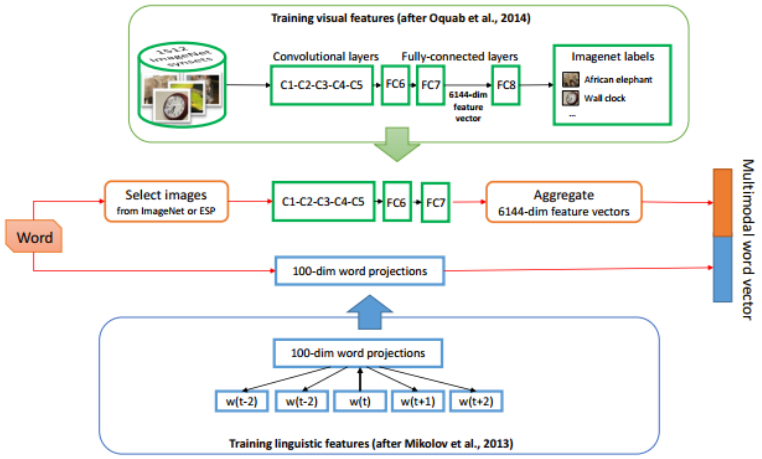

Figure 7 - Image and Text Fused Representation [50]

In [50] image embeddings and image features are combined in a multimodal word vector. The convolutional network, trained for classification tasks, is used as a feature extractor aggregating a 6144-dimensionality vector for each image. The skip-gram linguistic representation, 100 features, is computed by [51] for image embeddings. Those two outputs are later concatenated in a vector for joint semantics in a single space model.

Another approach [52] is to combine image fragments, extracted by an image CNN, with

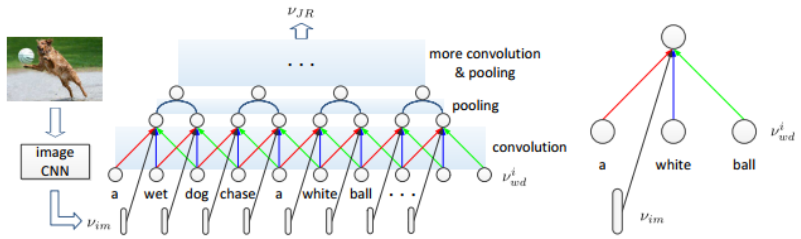

Figure 8 - Matching CNN Combined Representation [52]

sentence fragments in a multimodal matching CNN. The matching CNN's goal is to learn how to combine the two fragments producing a joint representation. Lastly, a Multilayer Perceptron (MLP) summarizes and evaluates the matching score.

Two-Stream CNN [53] applied for action recognition in video consists of two CNNs with inputs: a single frame and an optical flow multi-frame extracted from stopped frames

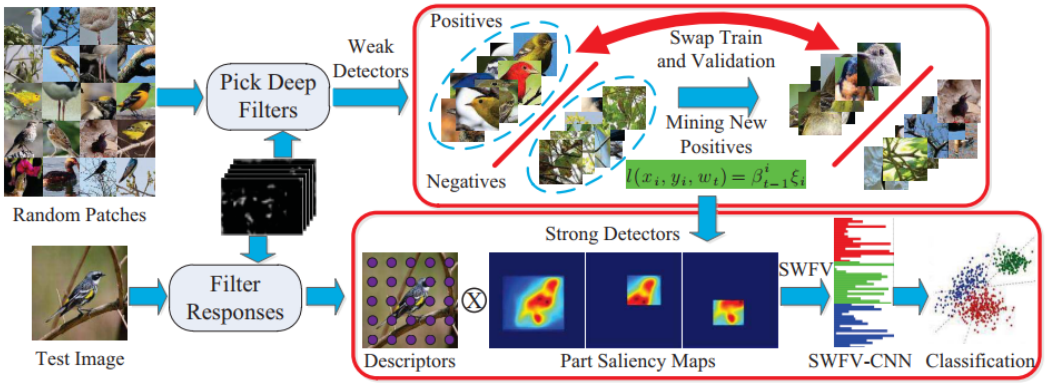

Figure 10 - Fisher Vector Combined with Convolutional Neural Network [54]

incorporating motion features. The class scores of the independent architectures are later fused, imitating the two-stream functionality of the human visual cortex, ventral and dorsal.

2-3-4. Hybrid Networks

Hybrid architectures combine "shallow" and deep algorithms for different tasks, like in Spatial Weight FV-CNN [54] for recognition, for image retrieval [55] and for texture classification [56].

Usually, they utilize a deep Convolutional Neural Network as a simple automated segmentation or feature extraction module trained from scratch with their own datasets or using a pre-trained model.

The "shallow" Fisher Vector pipeline transforms the incoming dense SIFT descriptors, computed from the subsampled segmented region, to a compact representation of 2Kd-dimensionality, K for the number of Gaussian components and d for the descriptors length. A

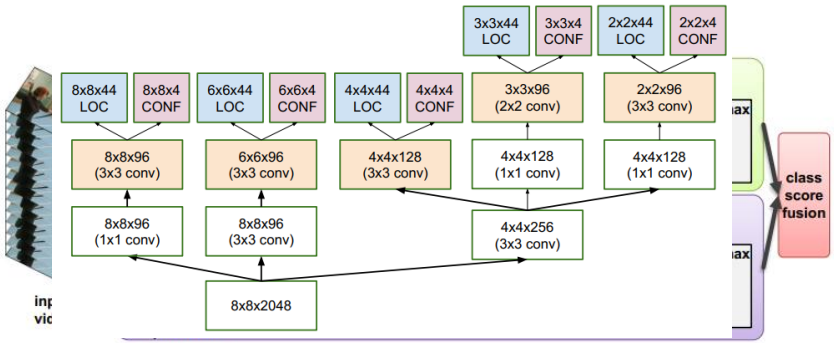

Figure Figure 11 - Bounding box prediction MultiBox [58]
Recog...

Support Vector Machine serves as the classifier of the extracted (CNN) and transformed (FV) image features.

2-3-5. Object Detection

DeepMultiBox [57] uses two deep CNNs, one for class-agnostic bounding box detection with 4 outputs for coordinates and 1 for confident, and the other for class labeling in an offline cascaded fashion. This implementation is an improvement over the original Multi Scale Convolutional MultiBox (MSC-MultiBox) [58] which follows a similar approach.

Subcategory detection [16] addresses the usual problems of detection related to object variation in scale, occlusion and truncation in everyday use cases similar to KITTI [59] or

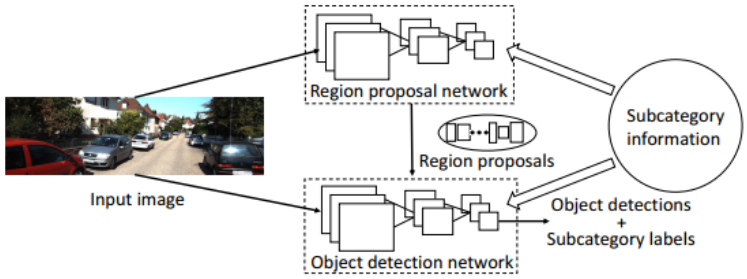

Figure 12 - Subcategory-aware CNN for Detection [16]

CamVid [60] datasets. This CNN architecture leverages multi-scale inputs, feature map extrapolation layer and a Region Proposal Network that outputs the bounding ROI regressor and its probability estimation.

The fast region proposal network (faster R-CNN) [61] is fully convolutional with full-image input for real time detection. Uses pre-trained convolutional models as feature extractors and a fully-connected module on feature maps for bounding box detection and classification. The detection speed is 5 to 17 frames per second using graphics processing units during inference.

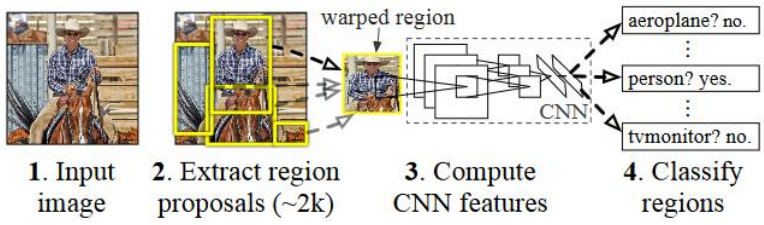

Figure 13 - R-CNN [62]

This is based on R-CNN [62] built on the same principles but utilizing AlexNet architecture. In [21], simultaneous detection and segmentation is achieved by combining feature vectors from a R-CNN and a CNN for bounding box predictions from hundreds of proposals.

VoxNet [17] performs real-time object recognition utilizing the powerful three-dimensional representation of volumetric Convolutional Neural Network on RGB-D natural images. A 3D CNN learns complex voxel filters leveraging the depth dimension, as a fundamental parameter of the model and resulting in richer feature maps with better understanding of the actual structure of the surrounding environment of the natural images.

2-3-6. Raw Feature Extraction

Pre-trained CNN models, even from a different domain, can be used as off-the-self feature extractors, like in [63], where the authors extract activation vectors from multiple scales of the same image, concatenate them and perform classification using a linear SVM evaluating the generalization property of those models. Similarly, a pre-trained OverFeat [64] used as feature extractor on different datasets, [65] [66], achieving competitive results and verifying the generalization property of deep architectures. Other approaches [67] combine activation vectors assisted by multiple pooling layers [68] or from various depths of the network for efficient

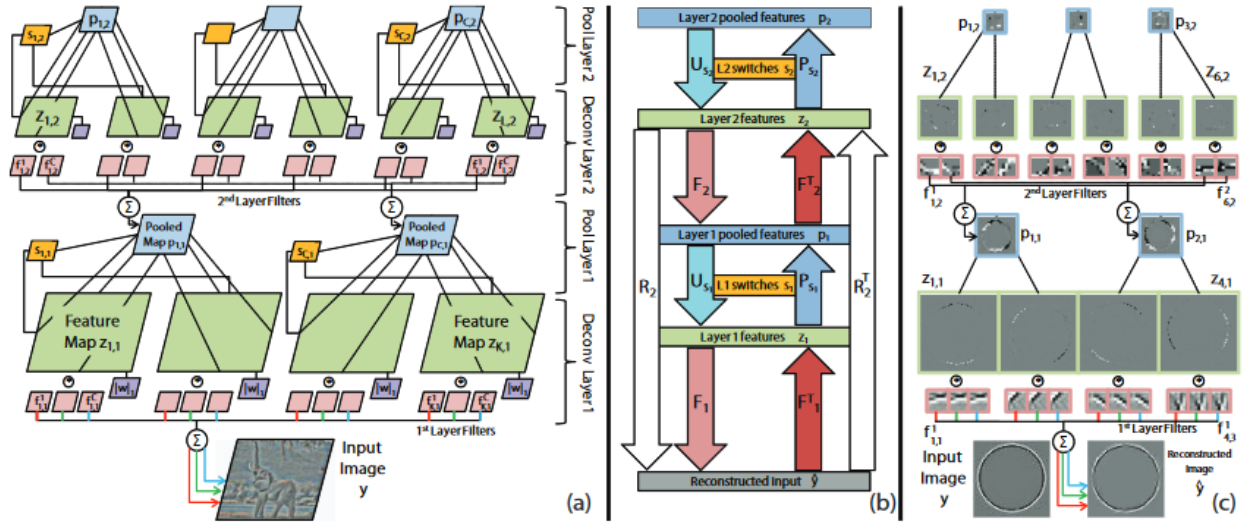

Figure 14 – Features [69]

texture recognition. Convolutional autoencoders [69] use unsupervised training for learning sparse coding and descriptive features from unlabeled data by backpropagating the reconstruction error.

2-3-7. Image enhancement

Image Enhancement by achieving super resolution [70] in context of DL, maps the low to high resolution images. Initially, the network encodes the low-resolution inputs, overlapping

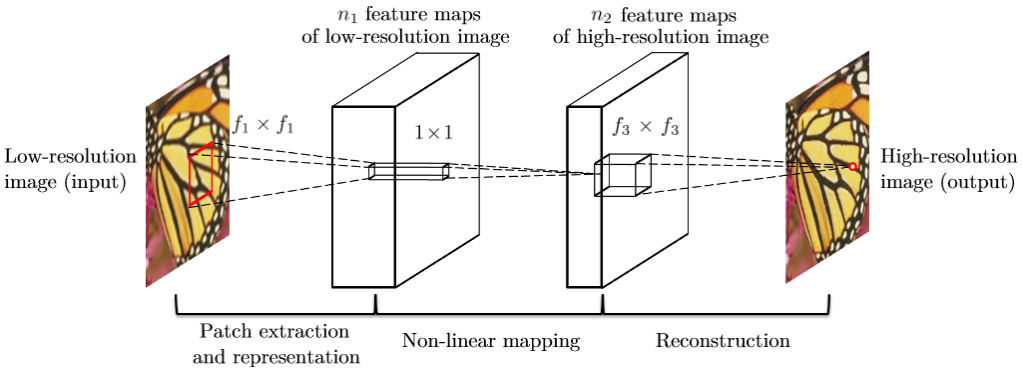

Figure 15 - Super-resolution [70]

dense patches, into a dictionary. The sparse coefficients are passed to the corresponding high-resolution dictionary, used for reconstruction of high-resolution patches. Finally, compose the output image from those patches. This sparse coding methodology is learnt by the convolution network in an end-to-end fashion.

2-4. Recurrent Neural Networks

Despite CNN being the state-of-the-art in image related tasks, a lot of researchers also experiment with the adaptation of Recurrent Neural Networks (RNNs) into computer vision. Previous successful paradigms of RNN include character prediction [71], machine translation [72] [73] or handwriting recognition [10].

Pixel-RNN [74] maps the distinct probability of pixel intensities and encodes every dependency in the image, resulting in a generative model that predicts the next pixel relying on all previous pixels and their non-linear behavior.

The ReNet [9] uses 4 RNNs to sweep images at every direction replacing a convolutional/pooling layer. Contrary to CNN, a major drawback for this type of architectures is

Figure 16 - Pixel-RNN test image completions

the inability of scaling computations in parallel due to the sequential units in RNNs leading to inefficient training times. ReNet can also be modified for semantic segmentation tasks. ReSeg [7] combines CNN feature maps with ReNet-based layers that captures both global and local spatial structure of the image. To match the dimensions of predicted segmentation mask to the original image an up-sampling block is introduced. For this task transpose convolutions are employed yielding resources and computational efficiency.

2-5. Unsupervised Models

2-5-1. Autoencoders (AE)

Autoencoders (AE) [75] are deep networks of fully-connected neural layers, with generative properties and trained with unlabeled data. The two-part setup consists of the encoder that produces a compact representation, code vector, of the input and the decoder that reconstructs the input from the code vector, learning, approximately, the identity function such as

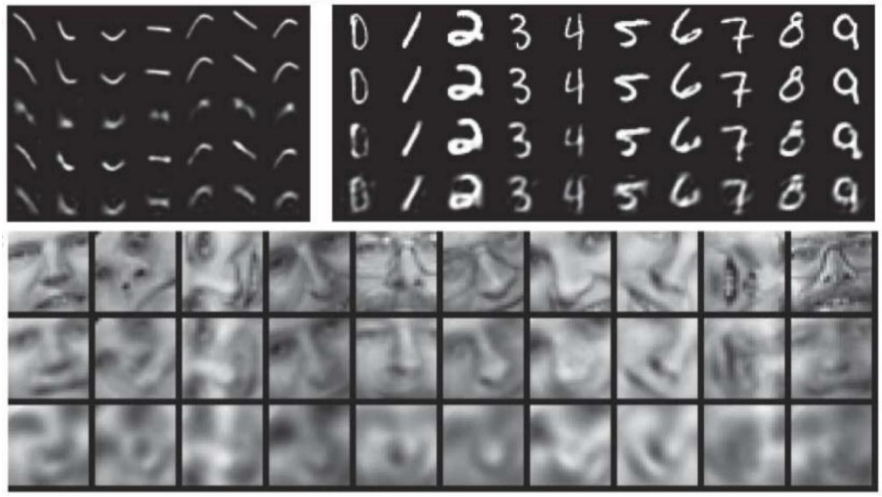

Figure 17 - Reconstruction of the original test images with autoencoders and PCA (top to bottom) [76]

the output is like the input. If an additional sparsity constrain is applied [3] then most of the neurons are inactive during training iterations enabling easier training and sufficient representations.

Hinton claims [76] that dimensionality reduction can be achieved by AEs, with their multiple adaptive layers, more efficiently than Principal Component Analysis, especially considering the reconstruction property of such representation.

Nonnegative Constrain Autoencoder (NCAE) [2] learn a sparse parts-based data representation due to nonnegative restrictions. When unsupervised training achieves a good reconstruction error, classification with deep stacked NCAE can also achieve competitive performance. Supervised fine-tuning with nonnegative weights and an additional softmax layer is required for such tasks. Stacked AE (SAE) [77] have been used for trajectory detection with transfer learning from an AE for feature extraction to the online tracker. Also, SAEs can be combined with convolutional layers forming the Convolutional Autoencoder stack (CAES) [78] serving as an unsupervised pre-training step to supervised CNNs.

Auto-Encoding Variational Bayes (AEVB) [79] have equal number of neurons for the generative model and the variational approximation part. Utilizes the Stochastic Gradient Variational Bayes (SGVB) approximator that can learn efficiently an inference model utilizing ancestral sampling.

2-5-2. Deep Belief Networks (DBN) & Restricted Boltzmann Machine (RBM)

RBMs [80] consist of constricted neurons forming a bipartite graph and are used in classification, dimensionality reduction and feature learning. DBN [81] is a generative model of stacked RBMs.

Convolutional Deep Belief Networks [4] [5] are a combination of convolutional blocks and DBN for hierarchical representation with scalable performance from benchmark datasets, like cifar-10, to larger images. The architecture's probabilistic inference achieves translation invariance from top-down and bottom-up.

2-5-3. Generative Adversarial Networks

The latest development in DL networks is the Generative Adversarial Networks (GANs) [22], based on the premise that a machine could better understand what is creating. The adversarial process includes two models: a generative G and a discriminative D trained simultaneously with the objective goal of model G to create believable enough artificial data, therefore the model D could classify them as real. This process is rooted to the game theory where an objective is achieved through competition and conflict of actions between two intelligent parties. GANs have been applied in data augmentation [82], realistic image generation [83], text-to-image synthesis [84] or even new drug generation in medicine [85].

2-6. Training Strategies

The representation of a highly varying function with few parameters can be achieved by combining many non-linear functions. Models with multiple consecutive layers manifest that behavior. The weakness of the deep architectures and with gradient-based optimization in general is that they start from randomly initialized weights with no incentive related to the problem they try to solve.

Thus, their initial value of weights during training is important for the adaptation and convergence to an optimal model. Unsupervised pre-training [86] contributes to the initialization of weights near a local minimum benefiting the supervised training of the network in a way that is observed [87] and mathematically sound [88].

The author in [89] proposes a greedy layer-wise pre-training process for deep networks, starting with one layer at a time in an unsupervised way and finally fine-tuning the whole network to the domain of the problem.

Transfer learning from a source model to a model of different domain is an alternative to

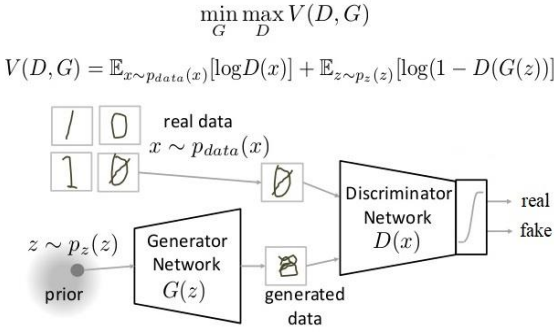

Figure 18 - Generative Adversarial Networks architecture [197]

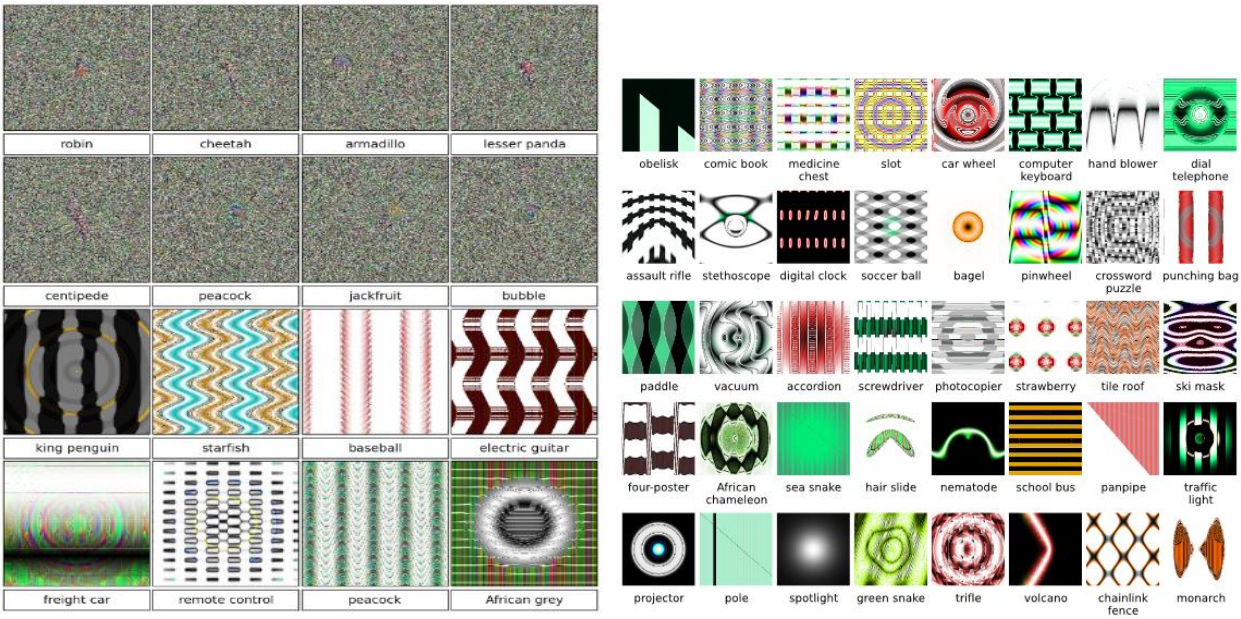

Figure 19 – Some predictions of unrecognizable images with very high confidence [92]

unsupervised pre-training. According to [90], an effective way to fine-tune a pre-trained model is to start from the deeper layers and if necessary, gradually, in a layer-wise fashion retrain the shallow ones until achieving the best performance.

2-7. Weaknesses of Deep Learning Algorithms

Regardless of the obvious data and resources related restrictions, machine learning models can be vulnerable to adversarial examples [91] and misclassify similar samples to the ones that classified correctly. Simple models with little optimization are the most vulnerable due to their inability of adaptation to the nonlinear high-dimensional space. This indicates that the model could not learn the underlying concept regarding each class. Rubbish class examples are commonly producing from those models. Another research [92] shows that it is possible to achieve high confidence with random generated, unrecognizable images using universally accepted models like LeNet and AlexNet.

2-8. Challenges & Publicly Available Datasets

One of the most difficult tasks in applying Deep Learning it is not the assembly of several layers, the training process or the hyperparameter optimization but the data collection and annotation in a robust, descriptive of the problem and unbiased database. Huge amount of high-quality and labeled data are required for efficiently trained and highly performing models. To further democratize and advance the AI research different organizations and research

groups have gathered and evaluate data to create benchmark datasets. Some of the most influential and widely used datasets and the corresponding challenges are presented in this section.

Pascal Visual Object Classes (VOC) [93] was an annual competition starting at 2005 until 2012 with 5 challenges: classification, detection, segmentation, action classification and person layout on a public dataset with ground truth.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [94] is another visual competition, since 2010, where researchers test their proposed algorithms on a subset of ImageNet dataset with 1000-class. AlexNet (error 16%) was introduced in ILSVRC 2012 kickstarting deep convolutional networks. Leading to 2017, most proposals in that ILSVRC achieved less than 5% error.

LabelMe [95] is a database with annotated large images with scenes and not just objects cropped in patches. It includes three types of annotations bounding boxes, polygons and segmentation masks.

The SUN RGB-D [96] dataset is composed by high-quality images with 3D annotations. The additional context from cameras with depth sensor is useful for scene understanding, semantic or instance segmentation [18].

Two of the most benchmarked databases for evaluating and comparing computer vision algorithms are the Modified National Institute of Standards and Technology database(MNIST) [97] with handwritten digits and the Canadian Institute for Advanced Research database (CIFAR) [23] with small natural images. Both consist of 10 classes with thousand images per class and are widely used by the research community when a new algorithm is proposed to demonstrate the potential performance.

The ImageNet [98] is an ambitious dataset including millions of high-resolution, annotated images collected using a cloud-based tool, the Amazon Mechanical Turk. The aim is to represent most of the thousands synsets included in the WordNet lexical database.

CamVid database [60] offers pixel-wise ground truth for 32 classes like cars, pedestrians, cyclists and other objects at three daytime lighting conditions captured by cameras on a moving car providing corresponding context. In a similar fashion, KITTI dataset [59] is purposefully aimed in autonomous driving with more than 6 hours of recorded 3D annotated, stereoscopic footage in high-resolution providing real-world traffic scenarios. The annotation bounding boxes follows the coordinated data of the 3D Velodyne scanner.

Brain Tumor image Segmentation (BRATS) [99] is a dataset with multimodal, annotated, 65 MR images of low and high-grade glioma, publicly available for the segmentation challenge associated with the Medical Image Computing and Computer Assisted Interventions (MICCAI) conferences.

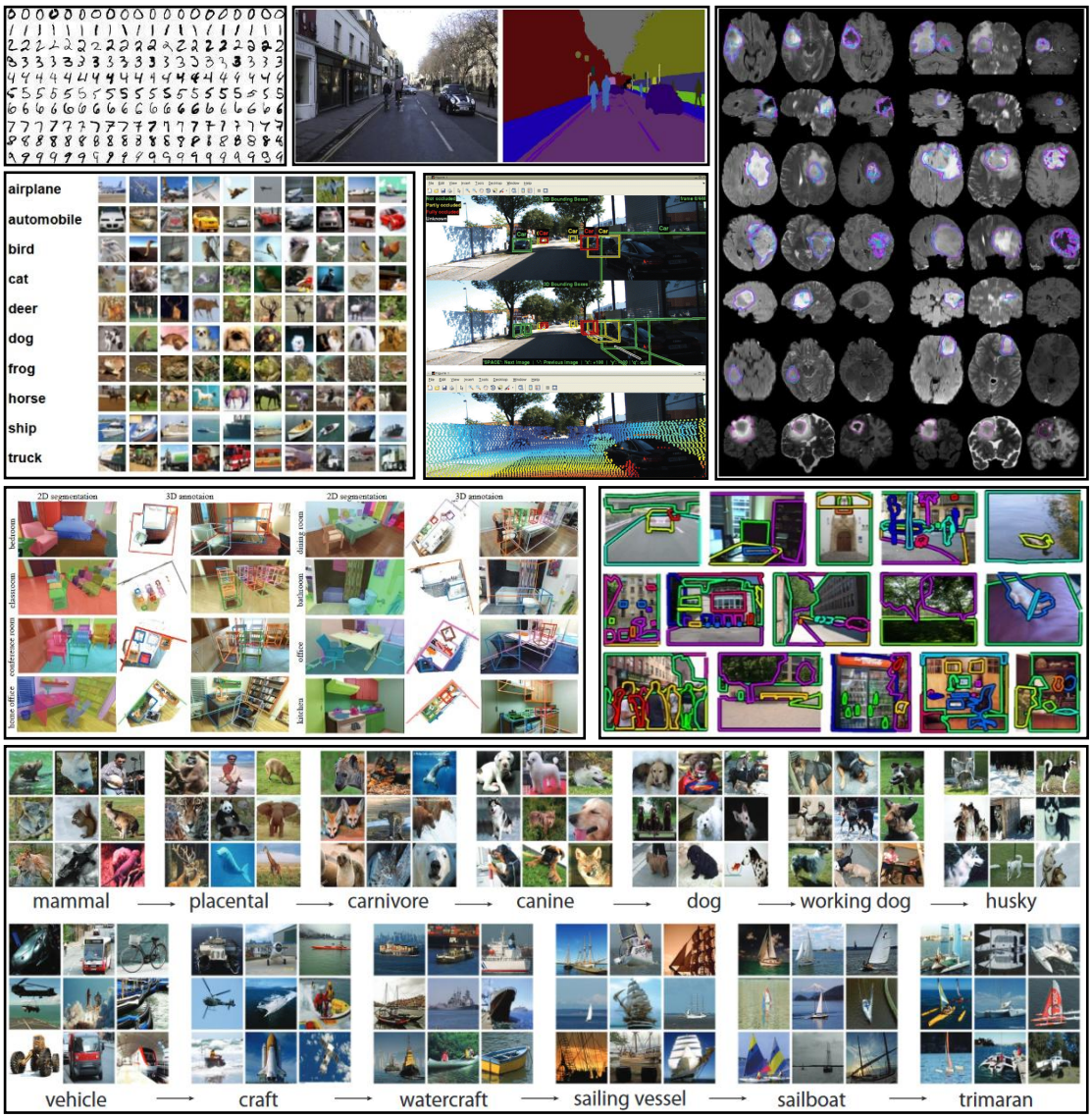

Figure 20 - Examples of databases MNIST [97], CamVid [60], CIFAR-10 [23], KITTI [59], BRATS [99], SUN RGB-D [96], LabelMe [95], ImageNet [98] (top left to bottom)

3. Medical Image Analysis

3-1. DL Applications on Medical Imaging

The past few years a shift of research trends in computer vision has occurred, from "shallow" or handcrafted algorithms to deep architectures based on learning. The massive state-of-the-art performance of the latter [36] [38] [44], the low-cost processing power and the availability of big annotated datasets [23] [60] [95] [96] [98] in various domains contributed to the rise of the new discipline of Deep Learning.

The biggest differentiating factor for the two schools of thought is the relationship of feature filters to the problem that try to solve. For instance a typical handcrafted algorithm like Fisher Vector pipeline [30] consists of densely extracted descriptors [28], dimensionality reduction [29], encoding [25] and finally classification or clustering. Each of these steps is a separate algorithm with different objective. It is also easy to observe that little knowledge about the problem is incorporated in the process itself.

On the contrary, the biggest advantage of deep architectures is the simultaneously training process of feature related tasks like extraction, selection, reduction and classification in a merged, end-to-end learning process. Additionally, the low-level feature filters are shaped by the dataset.

A few attempts have been made to combine the best of both worlds like in deep fisher kernel [33] where instead of separate steps the Fisher pipeline is merged into an end-to-end deep layer with objective to solve one convex optimization problem. Also, in [55] and [56] the authors propose hybrid architectures of Convolutional Neural Networks followed by Fisher Vector encoding for image retrieval and texture segmentation respectively. Compared to traditional methods some gains were observed but not state-of-the-art level performance. It is useful though to view them as an alternative in cases where the requirement for large datasets are not met.

Deep Learning architectures have also been transferred successfully into the medical domain with implementations for classification, segmentation, detection, registration, image retrieval, enhancement.

Particularly, for medical image analysis fine-tuning [90] [100] of pre-trained models tackle an important issue regarding the availability of adequate dataset and simultaneously performing equally or better than trained from scratch models.

3-1-1. Classification & Detection

DL classification of anatomical areas and lesions include: epithelial and stromal [101], breast cancer benign vs malignant [102], interstitial lung disease [103] [104], lung nodule with multi-scale CNNs [105], peri-fissural nodules [106] utilizing multiple pre-trained OverFeat [64] models, one for each view, as feature extractors and posterior probability fusion, calcium scoring with a patch-CNN [107] in cardiac CT images with encoded in color channel the tree views axial/sagittal/coronal, 3D U-net-like architecture [108] for CT scans auto labeling of vertebra centroids, Alzheimer's disease (AD) diagnosis using 3D CNN built on pretrained 3D CAE [109], similar 3D CNN for AD but with AE initialization [110] and retinopathy detection utilizing Inception-v3 architecture [111]. In [112] the author fuses the probability of three AlexNets for anatomy localization of organs in the 3D space with chest CTs acquired in axial, sagittal and coronal orientations.

The proposed prostate cancer detection in [113] combines U-Net as segmentation model S with purely adversarial training demonstrating improved sensitivity and DSC scores compared to other cross-entropy based training methods. The discriminator D, which is trained with expert segmentation masks, decides if the segmentation produced by S is correct or backpropagate the error until S minimize it and D maximize the accuracy of segmentation (DSC).

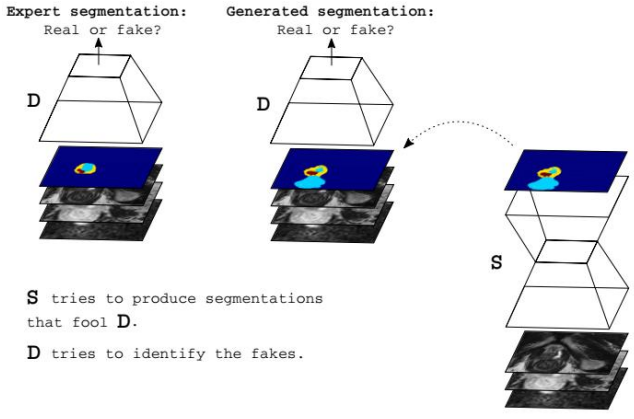

Figure 21 – Semantic Segmentation of prostate cancer utilizing GAN training [113]

Because of no standard anatomic orientation of lymph node, the recommended image

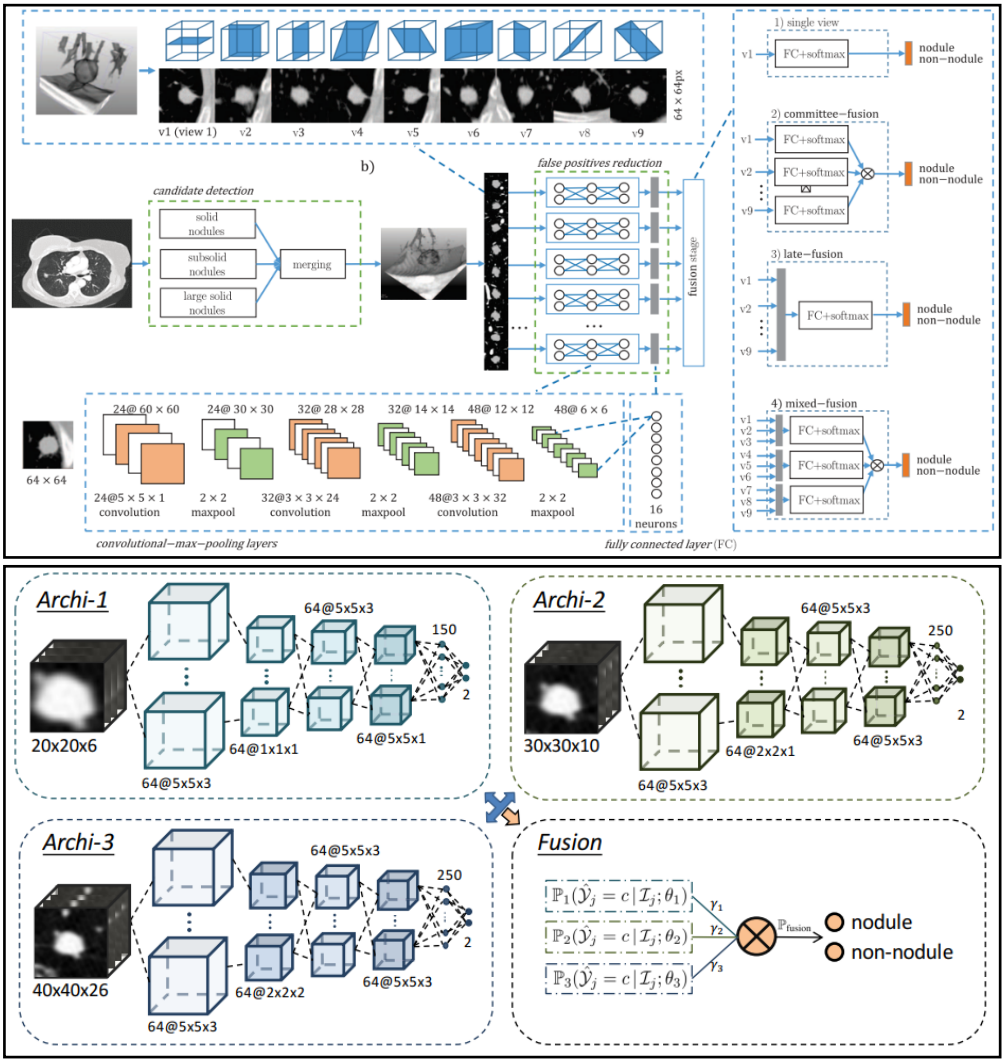

Figure 22 - Two different approaches to false positive reduction in nodule detection with multiple 2D [116] (top) and 3D [117] (bottom) CNNs

acquisition process is random. Meaning image resampling and reformatting can be done, like in [114] [115], by combining different views like axial, coronal, sagittal as three channels of information without spatial correlation among the channels. CNN model's representation does not have pixel-wise spatial correlation among input color channels. According to the author such transformation to RGB is allegedly compatible with transfer learning from pre-trained models with natural images. Another study [116] splits the detection into simpler networks and eventually evaluate their predictions by fusing into a robust universal prediction. Similarly, in [117] three 3D CNN, trained in different scale patches to capture the varying geometry for nodule detection from volumetric CT. The final prediction produced by fusing the posterior predictions of those networks.

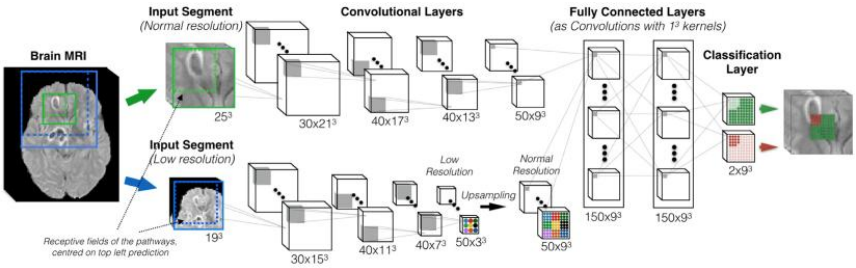

Figure 23 – Multiscale 3D CNN for brain tumor segmentation [131]

3-1-2. Segmentation Organ & Lesion

Deep learning segmentation have been applied in anatomical areas like breast [118], brain [119] [120], pancreas [121] [122] [123], prostate [124], cardiac [125] [126], knee [127], retinal tissue [128] by producing affinity graphs and organ localization [129] with modified FCN architecture. The different objective of deep segmentation architectures instead of other metrics dictates the use of Dice Similarity Coefficient [130] quantifying the similarity of output of an algorithm to the manually segmented image or mask.

The proposed architecture in [131] learns high-level and complex features separating Grey Matter, Cerebrospinal fluid and other anatomical structures and regions of lesions. Combines two-pathway 3D CNNs, for multiscale inputs replacing pooling layers, with a 3D extension of Conditional Random Field (CRF) as post-processing for the predicted segmentation masks.

Another implementation for infant brain tissue segmentation [132], specifically white matter, grey matter and cerebrospinal fluid, leverages the multimodal imaging T1, T2 and fractional anisotropy for better understanding the early stages of brain development in health. For

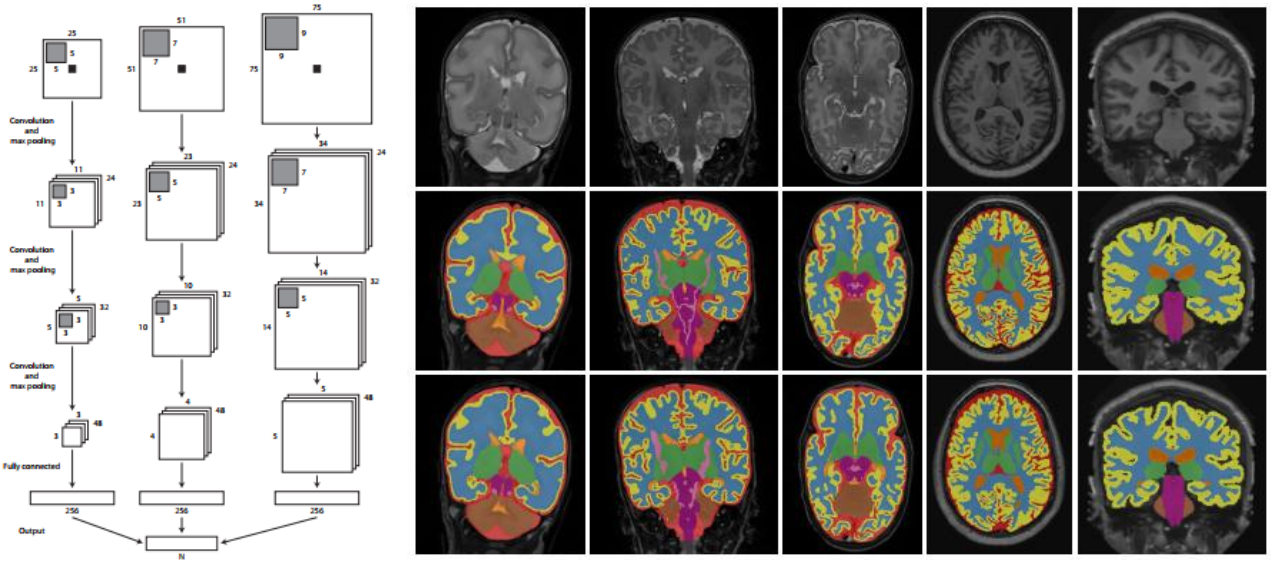

Figure 24 – Multiscale 2D architecture for brain segmentation (left) and segmentation manual vs predictions (right-middle vs right-bottom) [133]

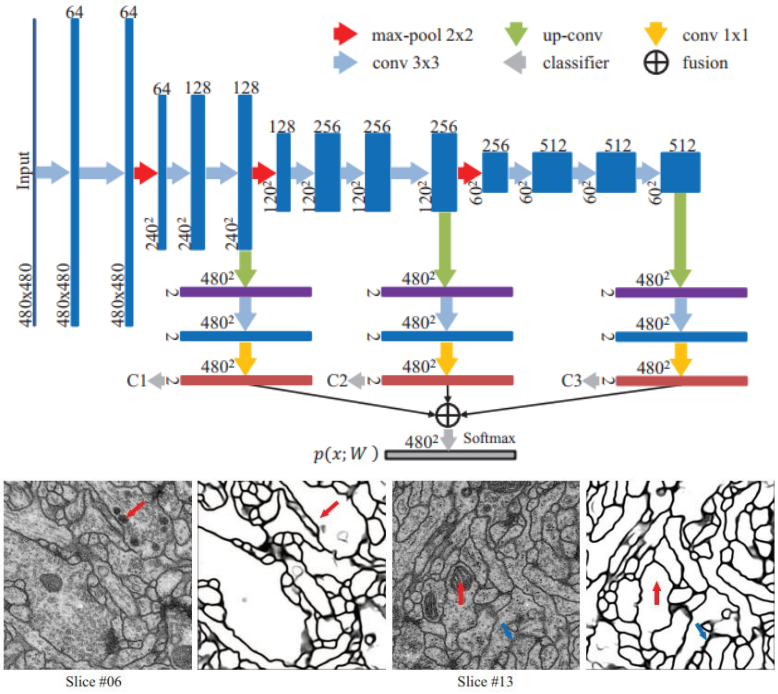

Figure 25 – FCN neural structure segmentation [138]. Architecture (top) and segmentation predictions (bottom).

the same domain, the author proposes [133] a more advanced 3-way 2D CNN with multi-scale patches to extract descriptive features from varying viewing perspectives. Other implementations for brain segmentation include: U-Net based [134], two-path CNN [135] and with the addition of CRF [136] for refinement.

Due to time and hardware limitations instead of using the demanding 3D CNN architecture the authors in [127] deployed three independent 2D CNN, one for each planar, with common only the softmax layer to segment the knee cartilage.

A popular architecture for segmentation is the Fully Convolutional Network [137] with skip connections, for combining activations from the shallower (finer) layers with deep (coarser) activations, resulting in more accurate predictions. Such architecture applied to datasets for neuronal structure segmentation [138] with the use of data augmentation. Combing two FCN models for cardiac segmentation [125] with transfer learning from the model of the left ventricle to the model of the right ventricle, the author deals with the lack of samples during training and simultaneously boosts the convergence of the model to a high-quality solution.

U-Net [139] is a type of Fully Convolution Network for segmentation specifically developed for medical images. Like any typical FCN utilizes an encoder-decoder architecture with connections that propagate feature maps from the encoder, convolutional, layers to the responding decoder, transpose convolutional, layers. A differentiating factor from SegNet is the

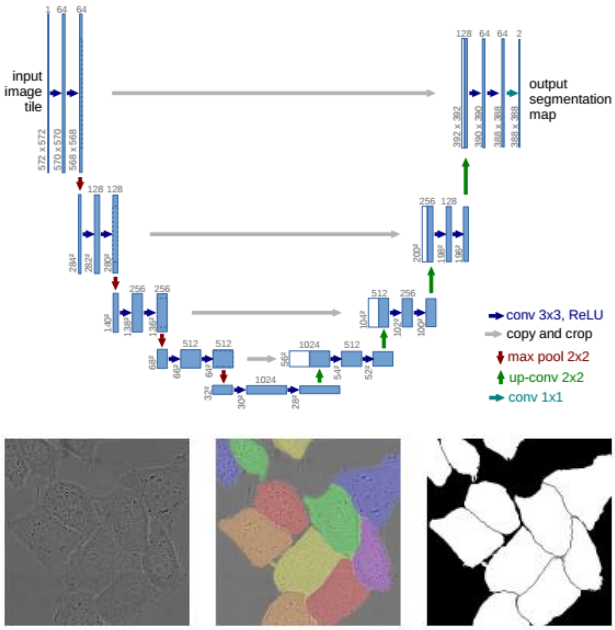

Figure 26 - U-Net architecture (top) [139]. Original image, manual vs predicted segmentation masks (bottom).

use of connections between these two parts demanding more resources in case of U-Net but yielding slightly better performance.

Most medical image data are spanning across three-dimensions as volumes. Thus, learning models that can take advantage of such 3D representations are important for accurate predictions, especially for segmentation which requires global understanding of the region of interest.

An extension to U-Net but with 3D operations is the proposed network in [140] allowing native support of volumetric data and sparse labeling. Similar principals applied to V-Net [141] with the differences being the activation function, the number of filters, pooling layers and the type of ground truth.

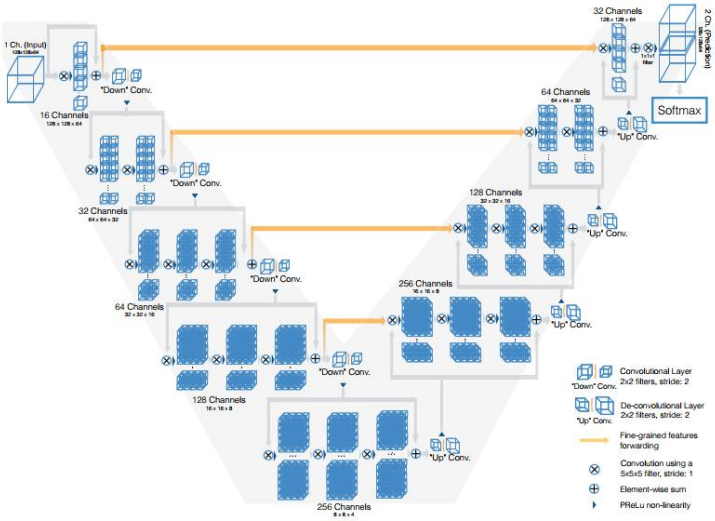

Figure 28 - V-net architecture [141]

Additionally, VoxResNet [142] along with volumetric segmentation employs residual connections between layers increasing the depth of network, showcasing competitive results in brain segmentation. Cascaded U-Nets followed by 3D Conditional Random Field [143] can offer a more robust lesion segmentation with 3D CT input data. In another direction, the authors in [144] proposed a cascaded architecture with parallel blocks of kU-Nets each capturing the intra-slice context of a 2D slice and two blocks of Bidirectional Convolutional LSTM, stacked in depth, to handle the sequential data from the patient's 3D scans. Another use of recurrent networks in [145] proposes a network with only Convolutional LSTM modules in different axes for 3D volumetric segmentation. Extending volumetric segmentation [146] with U-net architecture but instead of 3D convolutions each convolutional block consisting of four 2D convolutions of different viewpoints.

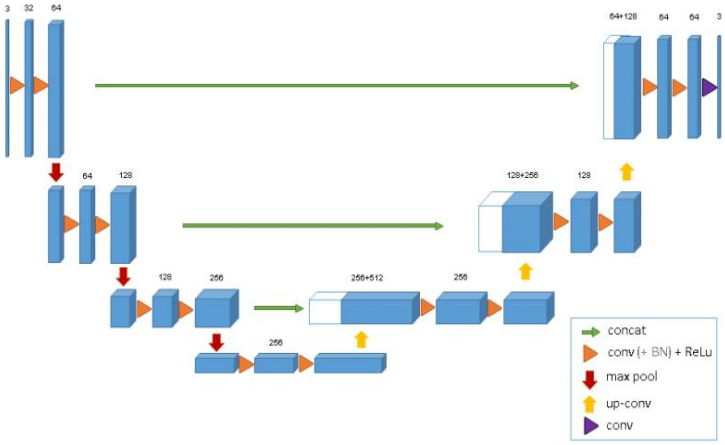

Figure 27 - 3D U-Net architecture [140]

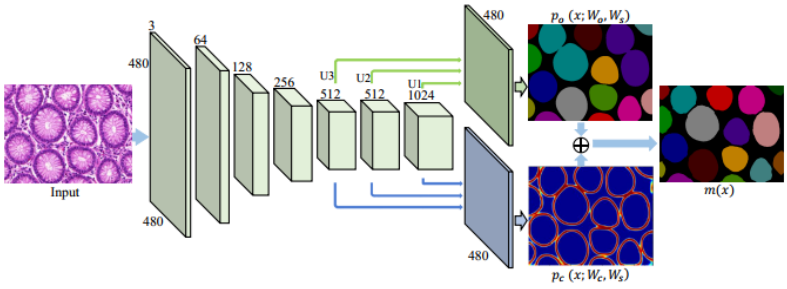

Figure 30 - Multi-level contextual features from FCN with domain adaptation for gland separation [149]

3-1-3. Transfer Learning

Transfer Learning to a target dataset with different distribution requires fine-tuning applied to the deeper layers that have learnt a high-level feature set representation related to the source dataset.

Indicatively, other architectures with weights transferred from related-domain or natural images trained models to: prenatal ultrasound segmentation [147], MRI brain lesion segmentation [148], two-path pixel-wise classification and contour-aware network [149] for

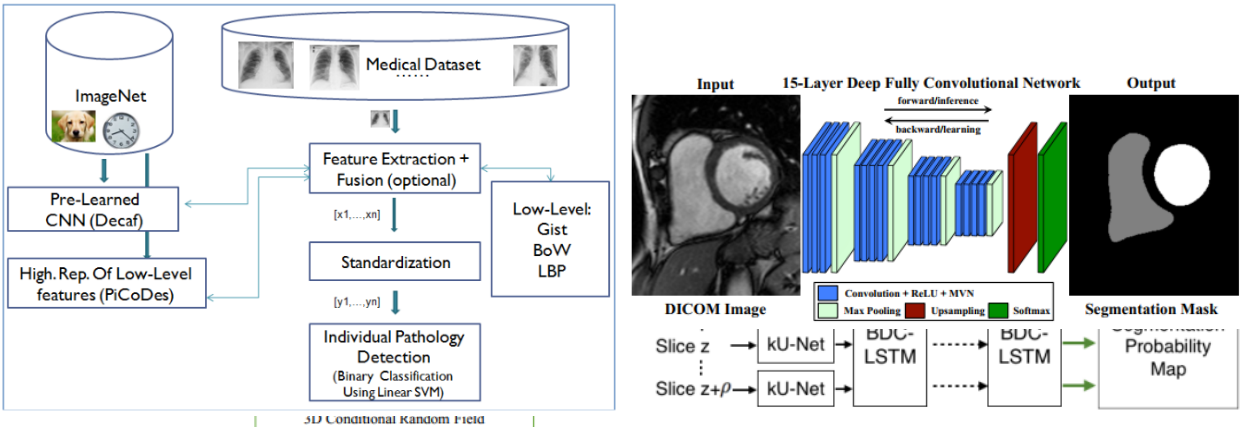

Figure 31 - Two transfer learning cases for chest lesion detection [150](left) and cardiac segmentation [125] (right)

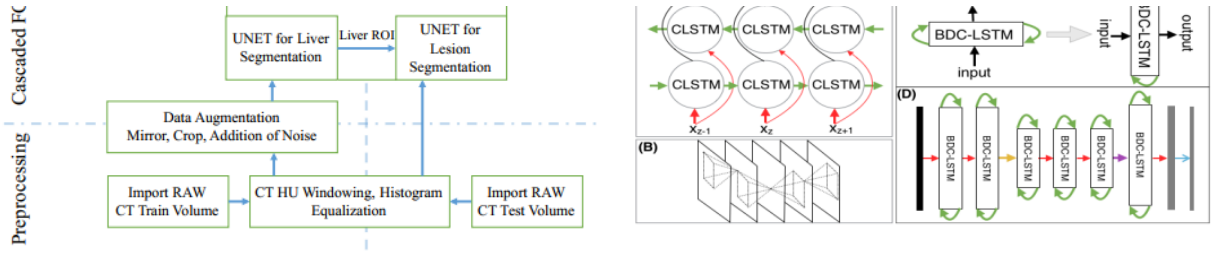

Figure 29 - U-Net combined with 3D-CRF [143] (left). kU-Nets BDC-LSTM architecture [144] (right).

gland separation, chest lesion detection [150], pancreas [121], cardiac [125].

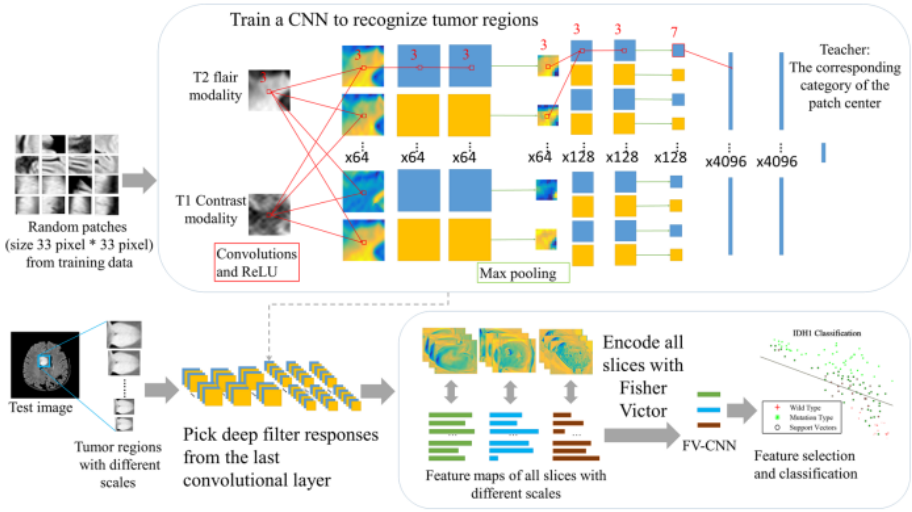

Figure 32 – Combination of CNN and Fisher Vector pipeline for discrimination of wild type vs mutation in low grade glioma [151]

3-1-4. Hybrid Networks

Hybrid networks, as explained in the Section 1-3-4, combine deep and "shallow" architectures, like in Deep Learning based Radiomics [151]. In this implementation the author utilizes Convolutional Neural Networks as a simple automated segmentation module, initially trained in a patch-wise manner to recognize tumor regions in patches.

Transferring the weights of the convolutional layers the author creates an only convolutional network, with a new input size, to segment unlabeled MRI images. Followed by a "shallow" layer for Fisher Vector transformation which includes step like subsampling, dense descriptors extraction, feature extraction and a linear SVM classifier [152] with the discriminating ability for the intraclass low grade glioma.

3-1-5. Registration

Image registration on medical data is important not only for diagnosis in general but also crucial to quantitative analysis of those types of data. To perform alignment of the fetal brain [153] on 3D ultrasound the author uses cascaded and multitask 2D combined with 3D convolutional networks. Initially the network decomposes the complex affine transformation, segment regions of interest and landmarks, map them into 3D space and learn the structural variability with specific coordinates.

3-2. Radiomics

DL techniques are not the first attempt to improve upon non-invasive diagnosis from medical images. In the early 2010s a new quantitative analysis of medical images was proposed. The combination of high-throughput computing and digital medical images provide high dimensional quantitative features, radiomics [154] [155], based on intensity, shape, size, texture and volume of tumor. The extracted features can be stored in large databases for future mining from multiple patients providing information on tumor phenotype.

The quantitative image analyses reveal information about pathophysiology and radiophenotypes of the examined lesion. Additionally, the correlation of radiomics with complementary data such as clinical or laboratory data can provide additive effect to Computer Aided Diagnosis systems. Considering that histopathology errors are up to 23%, non-invasive quantitative data can greatly improve traditional diagnosis.

As a result of such analyses, imaging biomarkers can be identified which can help cancer detection, diagnosis, prognosis, prediction of treatment response and monitoring of patient condition status.

Future advancements in the field will be highly dependent on the accessibility of cancer medical images and other medical data to researchers. Thus, it may enable the computation of vast amounts of radiomics, in big data scale, making possible the discovery of correlations among features and diseases. Those precise data are a requirement for better targeted treatments.

There are several key steps for producing radiomics, like image acquisition, delineation of volumes of interest (VOIs), segmentation of VOIs, extraction and qualification of descriptive features from volumes, feature storing to databases and finally information mining from those

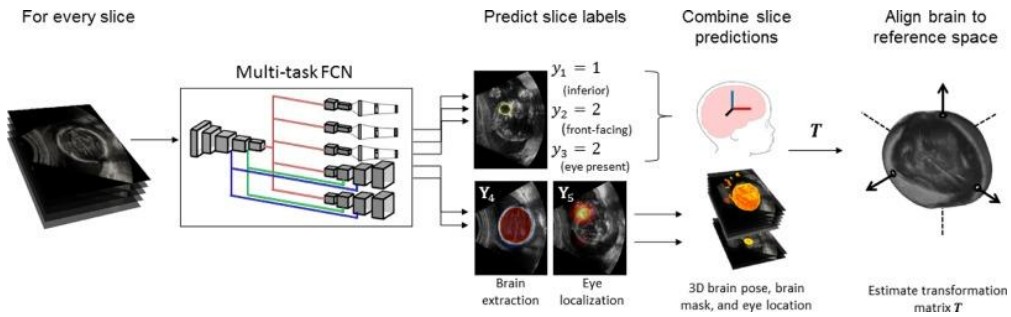

Figure 33 - Fetal brain alignment from ultrasound images [153]

databases.

Acquiring medical images follows standards and protocols, which varies among institutes. This is not important for clinical diagnosis but contrariwise, in quantitative level, it is difficult to compare medical images acquired by different protocols. Although, fusion of

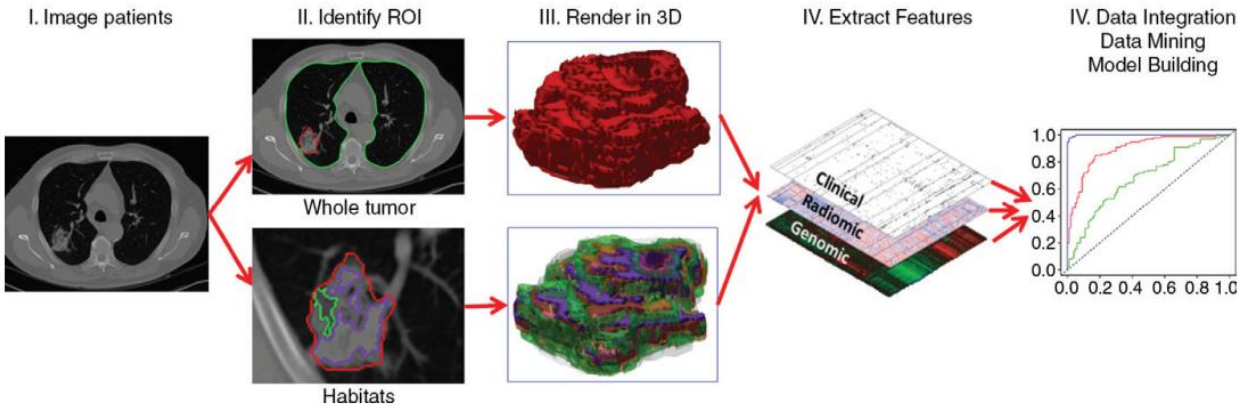

Figure 34 - The pipeline of radiomics [154]

different modalities like PET/MRI [156] can be used to provide multiple texture features.

The volume of interest delineations is extracted by combining multi-modal images for specific region designation, specifically tumor and habitat region. This is an important process for proper segmentation of those regions. Segmentation can be done by an expert manually or by specialized automatic algorithms and is a critical task for the computation of radiomics.

Feature extraction from high-dimensional data, like an image, is a result of statistical transformations of the semantic data, region of interest, providing the first, second and higher order features. It is worth mentioning that these descriptive image features can approach the complexity of gene profile expression.

Each order corresponds to a specific property of semantic data. The first order is histogram-based on voxels with no spatial correlations among them, the second order captures the statistical interrelationships or texture features and the higher order extract patterns from fractal analyses, Minkowski functionals, wavelets and laplacian of Gaussian transformations.

Ultimately, storing radiomics to databases is required for training effective predictive classifier models by discovering patterns and correlations among diseases and image features.

DL and radiomics pose similarities like the idea of extracting, transforming and selecting descriptive image features and finally process these massive amounts of data for creating models that enable personalized and targeted health-care. Differences are also evident especially the fact

that the radiomics workflow is not a unified process like in case of DL but rather multiple separate algorithms in a pipeline contributing added errors and biases at each step.

3-3. Review of DL approaches in Liver Cancer

Using DL for classification tasks researchers have deployed several implementations adapting state-of-the-art performing network architectures used on natural or real images that have a rich local covariance structure to the medical images that exhibit quantitative properties, and for this research specifically liver-related diseases.

Some examples in literature for liver lesions, including the use of histopathological images [157] for discrimination among different type of cancer stem cells in liver utilizing AlexNet and color deconvolution, liver lesions detection with patch-based CNN [158] and scale down input [159], cascaded residual CNN [160] using 2D CT slices, CNN with ultrasonic input for fat characterization [161], Broiler liver classification [162] with patches of high-resolution RGB images of poultry viscera captured, capsule for ultrasound image classification [163] of cirrhosis using pre-trained CNN, with MNIST dataset, as feature extractor and SVM as classifier

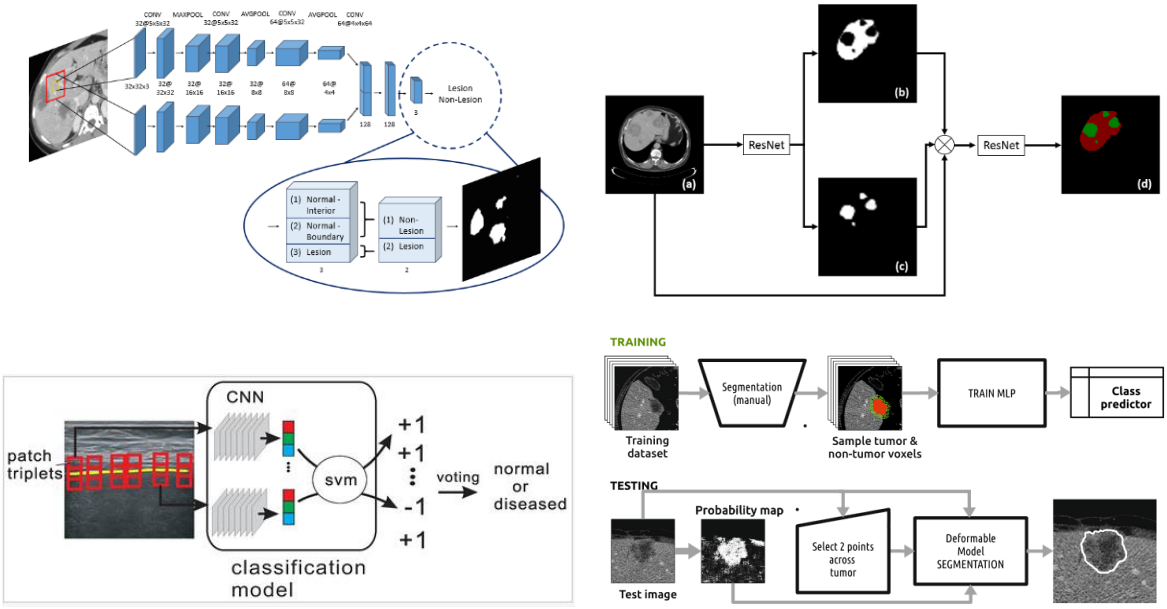

Figure 35 - Different architecture for liver lesion detection [158] [160], classification [163] and segmentation [164] (top-left to bottom-right)

or malignant tumor segmentation with MLP [164]. In [165] the proposed deep CNN utilizes wavelet-transformed low dose CTs as input to denoise the artifacts and reconstruct the images.

Computer Aided Diagnosis systems play a significant role in advancing the traditional medical field into precise, accurate and personalized health care. SurvivalNet [166] uses a

cascaded network architecture to predict the level of risk for the survival of the patient utilizing FCN for semantic segmentation and 3D CNN on diffusion MRI. Other decision support [167] uses CNN with CT scans.

Another important task for quantitative analysis or raw image classification is the segmentation either of the entire liver [168] or regions of interest [169] [170] using CT 2D patch-CNN, cascaded FCN with 3D CRF [171] using CT and the combination of two modalities CT/MRI [172], FCN [173] on Contrast Enhanced CT and with ultrasound images [174], FCN and lesion detection [175], using U-net for liver and dense U-net for lesion segmentation in a cascaded fashion [176] [177] and fully convolutional/deconvolution network [178].

The 3D Convolutional Neural Networks with volumetric inputs have been mainly used in segmentation tasks with CT volumes for liver lesions [179] [180], using 2.5D U-Net-like architecture [181] and densely connected 3D CNN [182], or on the whole liver [183] [184].

It is worth noting that there is a lack -at least to the best of our knowledge- of 3D CNN

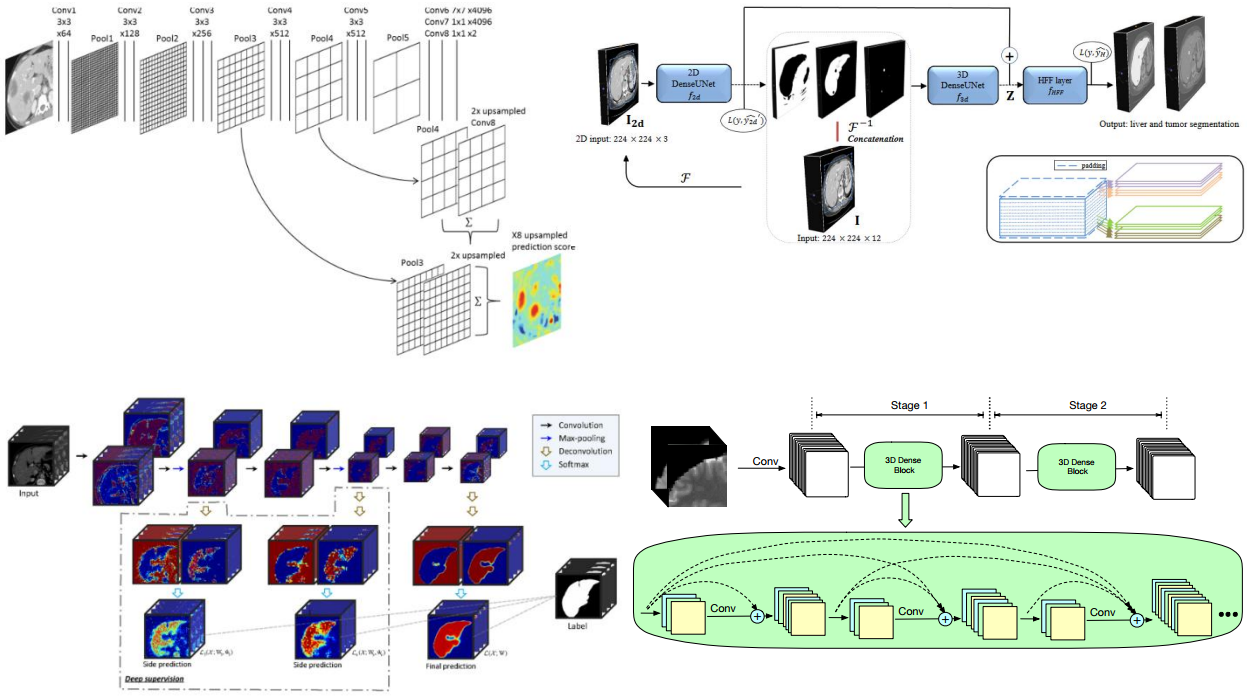

Figure 36 - Liver or lesion segmentation using networks like: FCN [175], Dense U-Nets [176], deeply supervised 3D CNN [179] and 3D Dense Blocks [182] (top-left to bottom-right)

architectures for classification or the use of 3D extracted features in liver cancer and lesion domain despite harnessing such representations in other domains like the natural image with RGB-D as discussed in Section 1-3-5. In the next chapter a comparison is presented between 2D and 3D convolutional networks in terms of discriminative feature extraction for tissue classification responding to the literature deficiency.

4. Convolutional Neural Network

Convolutional Neural Networks (CNN) extend the Artificial Neural Networks (ANN) to better handle image analysis tasks taking into consideration the spatial structure of the image in an end-to-end full-automated process. Currently, CNNs exhibit state-of-the-art performance in tasks such as classification, segmentation, registration and object detection used in robotics, autonomous vehicles, surveillance and bioinformatics.

CNNs was firstly introduced by Yann LeCun in 1989 [11] and used for handwritten numeric digit (MNIST dataset) classification achieving test error rate less than 1%. In the coming years following these advancements, deeper convolutional architectures like the AlexNet [38] managed to fully replace traditional image analysis algorithms and promote the Deep Learning concept in more complex real-world applications.

4-1. Architecture Inspired by Biology

Vision in biology is a system comprised of different organs that transforms photons of various wavelengths on retina into electrical signals transmitted to the primary visual cortex throught the optic nerve, as demonstrated in Figure 37.

Analogously, the CNN architecture uses convolutional, activation and pooling layers to transform the input image (millions of pixels) or any array into a compact representation

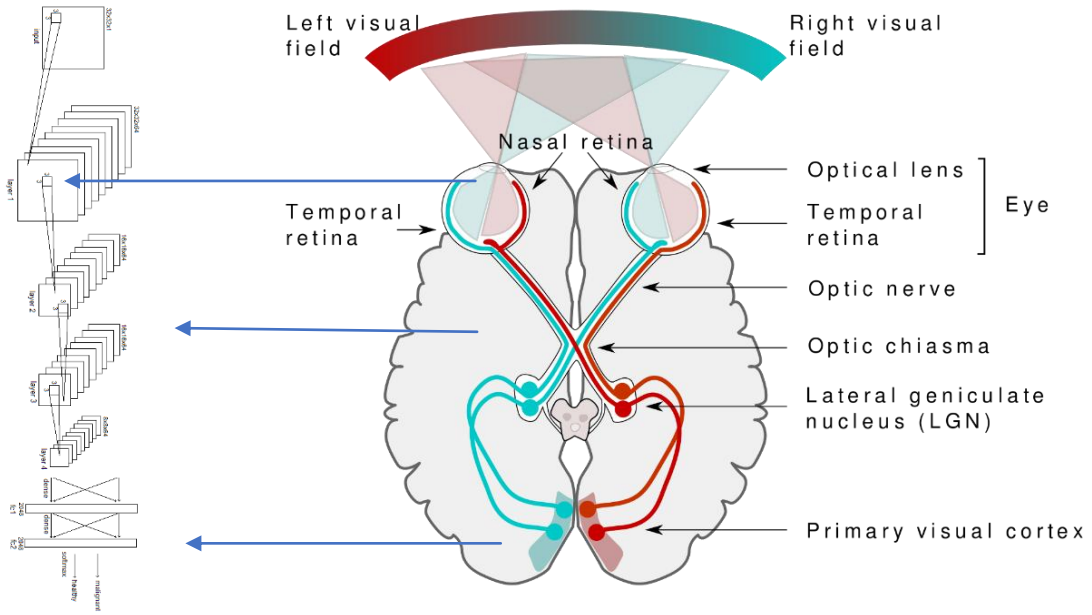

Figure 37 -Human vision inspired deep Convolutional Neural Network

(thousands of features) that is discriminative and sufficient for classification. Stacking many of these layers allows a hierarchical representation of visual information to be learnt introducing properties like scale, shape and translation invariance. In addition, deeper networks approximate better the non-linear relationships among the classes in the dataset.

4-2. Network Architecture

As mentioned earlier CNNs are comprised of many different layers including convolutional, batch-normalization, activation, pooling, neural and classification. This aims to improve the image analysis pipeline introducing one unified network with many consecutive layers with the only objective to minimize the prediction error by backpropagating the gradients.

4-2-1. Convolution

Convolutional layers calculate the dot product of learnt kernels and the sliding window area of input images with a predefined stride preserving the spatial information among pixels and capturing the local dependencies. Feature maps are the accumulation of these dot products over the whole input resulting in an alternative less detailed and more abstract version of the corresponding input. The dimensionality of feature maps depends on the padding, number of learnt kernels and stride or pooling.

Equation 1 - Convolution, w weights, x input, b bias

$$f(x) = w \circledast x + b$$

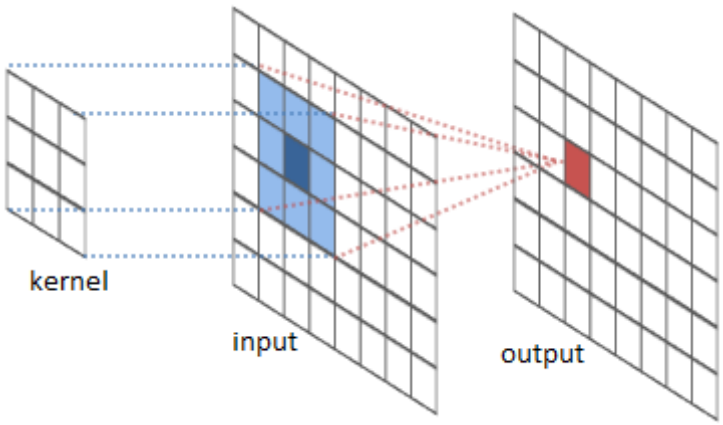

Figure 38 - Kernel convolving over input

Equation 2 – Number of Neurons n per layer, W volume size, K receptive field, P zero padding, S stride

$$n = \frac{W-K+2P}{S} + 1$$

4-2-2. Feature Detectors

Each convolutional layer learns multiple kernels to detect different features such as edges or curves. This enables the model to build high-level feature maps from low-level kernels by stacking multiple layers in a hierarchical input representation. Initially the network weights and kernel draw values from a gaussian, uniform or normal distribution in relation to the layer hyperparameters. During training the kernels adapt to better values through backpropagation of the gradients. The number of neurons for each convolutional layer can be computed by the Equation 2.

4-2-3. Activation

There are a few activation functions used in neural networks such as hyperbolic tangent, sigmoid, rectified linear unit (ReLU) along with their variations, depicted in Equation 3, Equation 5, Equation 4. The activation function is an element-wise operation that introduces non-linear relationships among the input values in a network. In the context of Convolutional Neural Networks, ReLU is the fastest and best performing activation function that reduces the vanishing gradients problem in deep networks.

Equation 3 - ReLu

$$f(x) = \max(0,x)$$

4-2-4. Pooling

Pooling is the spatial subsampling layer of a network that reduces the dimensionality of feature maps. Consequently, it reduces the number of learnt parameters and the computational complexity of the convolutional layers. As a result, the above improve the generalization capability and introduce scale and translation invariance of the model. Some types of subsampling include average, summation, max pooling or an increased stride in the convolutional layer.

4-2-5. Normalization

Batch Normalization aims to reduces covariance shift [185] of inputs by rescaling the mini-batch at the batch axis to values that follow unit variance and zero mean. Additionally, each Batch Normalization layer learns with gradient descent two parameters: gamma γ and beta β. Those restore the original activations if improvement of the expressive power of the network is needed.

Equation 5 - Hyperbolic Tangent

$$f(x) = \tanh(x)$$

Equation 4 - Sigmoid

$$f(x) = \frac{1}{1+e^{-x}}$$

Empirically other improvements include faster convergence of the model with higher learning rates, reduced overfitting due to the introduction of noise at the batch level from this process and increased stability of the propagated gradients.

4-2-6. Neural Network

The Neural Network (ANN) part in the Convolutional Networks (CNNs) is comprised of many hidden layers and a classification layer. In the hidden layers every neuron is fully-connected to all neurons of the successive layer. The classification layer has different activation function than the rest of CNN usually Softmax or Sigmoid. The input of the ANN part is the flatten output (feature vector) of the convolutional part of the network. Its main objective is to

Equation 6 - Total Error, p probability

$$T.E. = \sum \frac{1}{2}(target\ p\ -\ prediction\ p)^2$$

reduce, select and combine the incoming image features in an automated, end-to-end way in

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Figure 39 - Batch Normalization applied to x over a mini-batch [185]

order to match the non-linear correlations between the input and the predicted class.

4-2-7. Training Process

The training process uses the training set to converge the network by adjusting the weights in proportion to their contribution to the total error in order to learn to classify properly the testing set samples. The aim is to provide models that generalize and not memorize the dataset.

At the beginning all the weights are initialized, forward propagate the training samples, calculate the prediction probabilities of each class, compute the total error, calculate the

gradients of the error of all weights in the network by backpropagation algorithm and finally update the weights with gradient descent. The objective of this process is to minimize the prediction error of the validation set and is repeated until the model converges to a good solution, local or global minimum.

5. Thesis Research: CNN on Liver Cancer

5-1. Reasoning

The aim of this research is to implement a deep architecture for extracting discriminative features from malignant liver tissue, determine the quality of predictions for primary vs metastatic in tissue classification and integrate them in a broader Computer Aided Diagnosis system in the future. This intraclass problem is difficult or impossible, even for human experts, to make a concrete diagnosis with only non-invasive clinical data, Section 4-2.

The Convolutional Neural Networks are the current state-of-the-art in image related tasks like classification, detection, segmentation and even generation. To achieve such an outstanding performance, with regards to training, large amounts of labeled images are required.

Medical image datasets lack [186] the quantity and qualitative characteristics due to facts like the legislative framework that protects the anonymity of the patients, different image acquisition protocols and devices, high cost of data preparation or the large number of lesions fragmenting the datasets. Typically, the number of samples for those datasets are low for both training and testing purposes in the context of deep learning approaches compared to natural image datasets, Section 1-8.

To overcome such difficulties, taking advantage of other properties of medical data like volumetric DW-MRI, creative techniques and algorithms have been developed, artificially augmentation of samples, multimodal data utilization, hybrid heterogeneous architectures that combine the benefits of deep and “shallow” pipelines, transfer learning from different domains and training processes like patch-wise training and unsupervised pretraining.

5-2. Dataset & Clinical Analysis

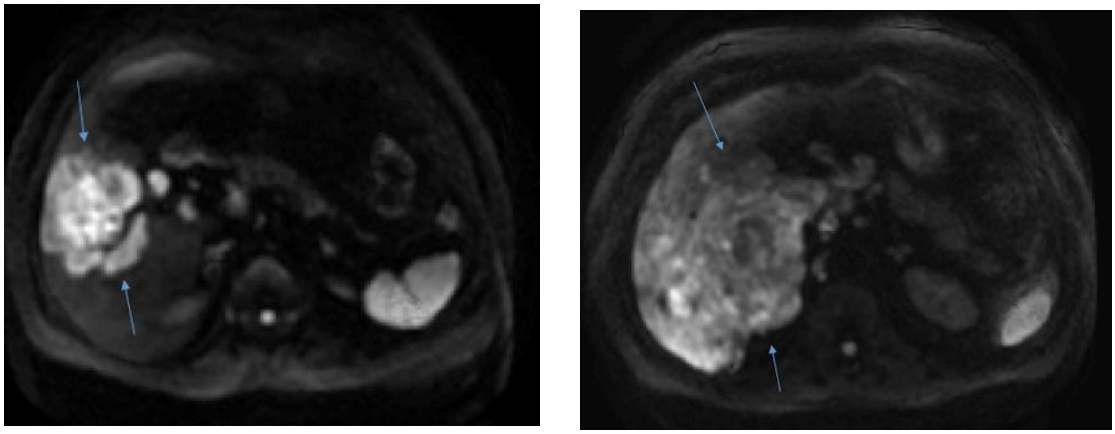

Figure 40 - High signal intensity on b 1000: Large Focal Liver Metastasis from pancreas (left), Primary Liver Hepatocellular Carcinoma (right)

For this study, 107 patients referred for upper abdomen MRI (mean age: 62 yrs., 17-87 yrs., 48 women and 59 men) were imaged at a 3 Tesla imager (Signa HDxT; GE Medical Systems, Milwaukee, WI) with twin-speed gradient system of 50 mT/m gradient amplitude. Data used for this study comprised primary liver malignancy (HepatoCellular Carcinoma: 21 patients, Cholangiocarcinomas: 16, or Hepatic AdenoCarcinoma: 3 patients) or secondary liver lesions of variable primary tumor site (breast: 11, colon: 10, lung: 11, pancreas: 10, neuroendocrine tumors: 8, prostate: 6, melanoma: 6, bladder: 2, stomach: 1, ovarian: 1, thyroid: 1) summing up to 107 patients with either solitary or multiple foci.

All images were obtained with an 8-channel torso phased array coil. A Single Shot Echo Planar Imaging diffusion sequence (DW-EPI) was acquired in combination with Parallel imaging (TR/ TE: 10.000/68 ms, slice thickness/ gap between slices 6 mm/1 mm, matrix 128 (phase) x 128 (frequency), field of view 400 mm, number of excitations 4 and receiver Bandwidth 1953 Hz, ASSET factor: 2). The total acquisition time ranged between 4 and 5 minutes due to the variation of respiratory patterns between the patients. Tumor delineations used for the 2D CNN and Signal to Error (SNR) measurements to assure image quality were calculated using in-house developed software.

Another key factor related to this type of cancer is the high variance in shapes and spread area, as displayed in Figure 40, where regions of interest can manifest as compact tumors or dispersed across a larger part of liver. Also, discriminating between primary or metastatic malignant entities based on tomographic images is difficult since they are both characterized by decreased extracellular space and therefore exhibit slower signal decay at high b-values. Thus, addressing the clinical problem of the liver tumor classification only b-1000 DW-MRI data were used.

5-3. Preparation of Dataset

5-3-1. Patch-Extraction for 2D CNN

The patches for the 2D CNN training were extracted from the dataset and separated into two classes: malignant and healthy tissue. The patches with malignant tissue include at least 40% pixels from the region of interest in order to achieve good recognition performance between classes. This offline process utilizes the GPU-powered Tensorflow [187] framework in a brute-force sliding window approach where every tumor slice is scanned by a 32x32 window with a 16-pixel stride.

Another factor that dictates these dimensions is the problem itself as the cancer region can manifest in a concentrated, compact area or with small dispersed regions of interest. Thus, using bigger patches will result in losing those small regions.

Crucial for this task is the delineation of accurate ground truth from either a human expert or by an automated algorithm. For this research the pixelwise annotations, regions of interest, were provided by an expert radiologist. In total, near 2,000 patches were extracted balanced from either healthy or malignant tissue.

5-3-2. Sample Selection

The input of the 3D CNN is the actual volume of the patient. The number of slices in each volume varies in our dataset from 20 to 80 slices but most samples have 30 slices. Considering that the input cannot be an arbitrary size in any CNN model some sort of selection was necessary.

Due to that requirement the samples with less than 30 slices were discarded completely. For samples with more slices I just discarded the extra slices with no regions of interest until the requirement was met. In one case the number of slices with ROIs was over 30, therefore I dropped some of them. Finally, 107 out of 111 samples were used for both training set and test set in both 3D and 2D architectures. No other clinical criterion for sample selection was applied.

5-3-3. Data augmentation

Data augmentation [188] [189] [190] is achieved by simple image processing techniques like deformation, multiple angles rotation, mirroring, scaling, brightness tuning and flipping. Particularly in this implementation, affine deformation with 9% alpha affine, 90^{o} and 270^{o} rotation, flipping up-down and left-right, brightness reduction and mirroring with 8-pixel wide reflections. Visual demonstration is depicted in Figure 41. This procedure managed to increase

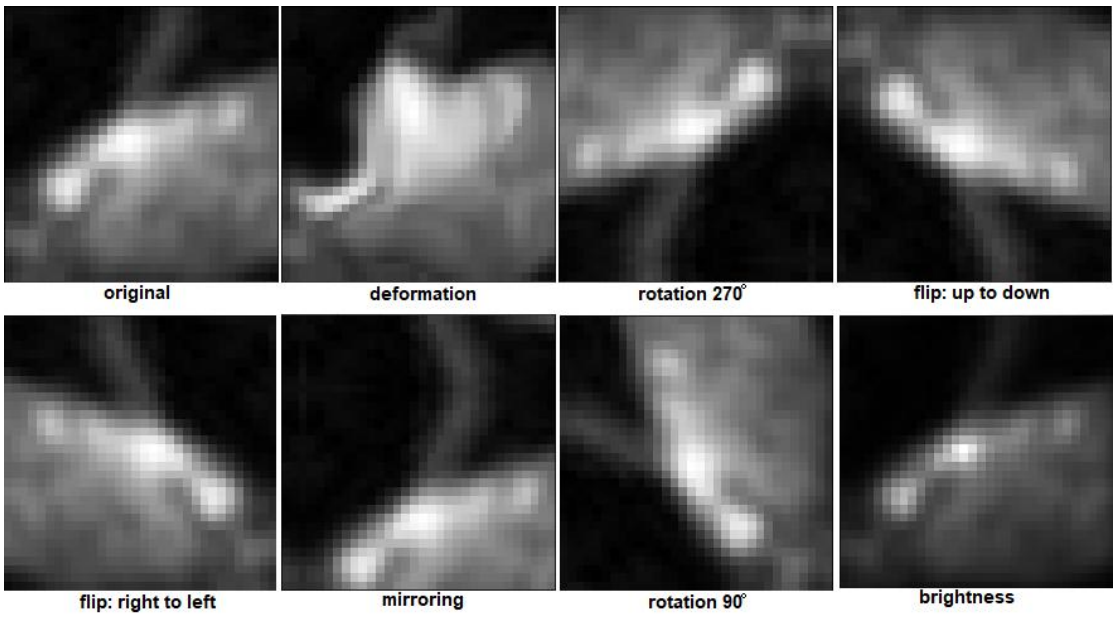

Figure 41 - Different types of data augmentation used in this research

the data by a factor of 10 resulting in over 20 thousand patches in total for the dataset of the patch-wise 2D CNN. Thus, augmentation was applied only to patches after the extraction from the sample slices. Besides, the increase of the sample count, data augmentation results in the translation, viewpoint, scale invariances [12] and artificial diversity of the examined dataset, thus strengthening the generalization capability of the model.

The data augmentation is necessary not only for inflating the dataset with more samples but also alleviates, to an extent, the presence of class imbalance. Specifically, this is a noticeable problem in case of 3D CNN where the dominant class in the dataset have around 40% more samples than the other class. After this process the larger class increased 7 times and the smaller by 10 times for the dataset of the 3D network and applied directly to the sample volumes.

5-4. Proposed Network Architectures

Convolutional Neural Networks [11] belong to a class of models that can learn hierarchical features by building high-level features from low level ones. CNNs adapt their internal representation through weight updating inside each layer by backward propagation of the calculated error from the loss function. Two main architectures were implemented and evaluated: a 2D CNN with patch-wise training and a 3D CNN trained on the whole volume.

5-4-1. 2D CNN and Feature Extraction from Slice

Patch-wise training is selected as a viable option due to the size of dataset, as it was discussed earlier in Section 4-3-1. Therefore, the architecture of this network includes 9 layers in

total: with input size 32x32, 4 convolutional layers with ReLU activation [191], 64 kernels and receptive field 3x3. A detailed description for the layer configuration is provided in Table 1.

Simplifying the architecture [192], no max-pooling layers are implemented, instead subsampling is achieved with strided convolutions with every second convolutional layer followed by batch-normalization [185].

Stacking layers in such fashion and not using wider layers and shallower network enable modular feature learning to avoid memorization by the network and achieve good generalization performance. For comparison a 4-convolutional with 64 filters per layer and batch-normalization performs the same with an 8-convolutional and 64 filters per layer but the latter is more prone to overfit and is harder to train.

The fully-connected part of the network consists of two layers with 2048 neurons each and 50% probability of dropout [193]. Glorot initialization [194] and Adam optimizer [195] with sparse cross entropy as loss function are selected. Finally, softmax layer serves as classifier with 2 classes for patches containing healthy or malignant tissue. The overall architecture is displayed in Figure 42.

After the patch-wise training, the produced model is able to discriminate among healthy and malignant tissue patches with accuracy over 97% and precision 92%. In other words, the feature extractor part of the network gives activations that enable the amplification of regions of interest. Thus, with transfer learning to a model with new input and new classifier could possibly discriminate between metastatic or primary malignant tissue from the whole image. But to extract features from a sequence of normal MR images, in this case 256x256x1, a new input must be appended resulting in the rejection of the fully-connected layer. The weights and layers of the convolutional part were transferred with no changes since they can discriminate between malignant over healthy tissue. The feature vectors, extracted from the new model, were labeled by the clinical results in two classes: primary and metastatic. Supervised training applied to SVM

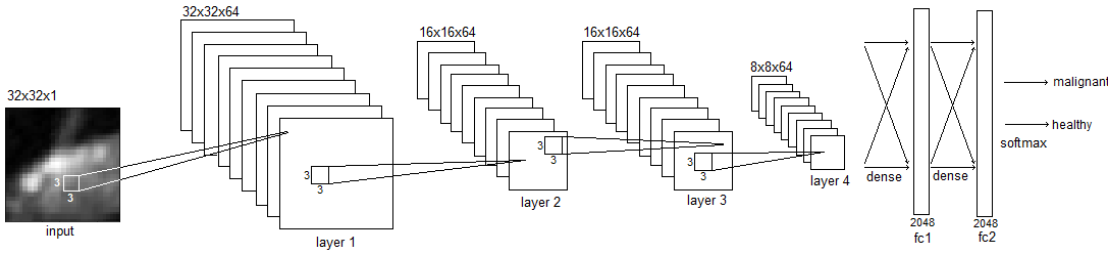

Figure 42 - Patch-wise training of the 2D CNN

Layer	Input	Configuration	Number of Neurons
2DCNN_1	32x32x64	2D Stride 2, ReLU	64x3x3
2DCNN_2	32x32x64	2D Stride 2, ReLU	64x3x3
Batch Normalization	16x16x64	-	-
2DCNN_3	16x16x64	2D Stride 2, ReLU	64x3x3
2DCNN_4	16x16x64	2D Stride 2, ReLU	64x3x3
Batch Normalization	8x8x64	-	-
fc1	4096	Dropout 50%	2048
fc2	2048	Dropout 50%	2048
softmax	2048	2 classes	-

Table 1 - Detailed layer setup for Patch-2D CNN

classifier corresponding samples like in CNN training.

Dimensionality reduction and normalization to the feature vectors also attempted with processes like Principal Component Analysis, thresholding feature removal or L2 normalization resulting in worse performance in classification. Therefore, I avoided any hand-crafted intervention.

5-4-2. 3D CNN and Volumetric Feature Extraction

The 3D Convolutional models learn more complex features spanning across three dimensions: width, height and depth. These relationships, at voxel-level, start from consecutive slices in early layers to more distant slices deeper, giving global knowledge about the internal tumor structure but also the overall geometry of the tumor respectively. This is the biggest advantage of that type of networks. Voxel-based information is already used with the 3D variance of radiomics [196] providing richer features.

Batch-normalization [185] is implemented after every ReLU activation of the network, including in total 4 strided [192] 3D convolutional layers and one fully-connected layer with 2048 neurons with ReLU [191] activation function. Filter size was set at 3x3x3 initialized by Glorot [194] and Adam optimizer [195] with learning rate 10^{-3}, 50% probability of dropout [193] was selected for better generalization along with a softmax classifier of two classes metastatic or primary. A detailed layer setup can be found in 2 and a visual representation of the network below in Figure 43.

Layer	Input	Configuration	Number of Neurons
Batch Normalization	30x256x256x1	-	-
3DCNN_1	30x256x256x1	3D Stride 2, ReLU	20x3x3x3
3DCNN_2	15x128x128x20	3D Stride 2, ReLU	20x3x3x3
3DCNN_3	8x64x64x20	3D Stride 2, ReLU	20x3x3x3
3DCNN_4	4x32x32x20	3D Stride 2, ReLU	20x3x3x3
fc	10,240	Dropout 50%	2048
softmax	2048	2 classes	-

Table 2 - Detailed layer setup for 3D CNN

Similar principles have been applied when designing both 2D and 3D architectures. The main differences are that in the 3D network the input samples are the DW-MRI volume, 30x256x256, per patient labeled by the clinical result, primary or metastatic. Consequently, there is no need for majority vote in evaluation like in case of 2D networks resulting in a process simple and straight forward.

Also, the extraction of features from the 3D network was correspondingly simpler, as displayed in Figure 44. The feature vector is the actual output of the fully-connected layer. The dimensionality reduction and feature selection were part of the training process and done by the neural layer. Samples for the supervised SVM training was selected, like in case of the 2D network, the same as 3D CNN training set.

5-4-3. Hyperparameters Optimization

The main parameters like the depth of network, number of layers, or the width of each layer, the number of filters and neurons of fully-connected layer were determined by a reasonably exhaustive research. The goal was the maximization of accuracy and the avoidance of overfitting and memorization by the network.

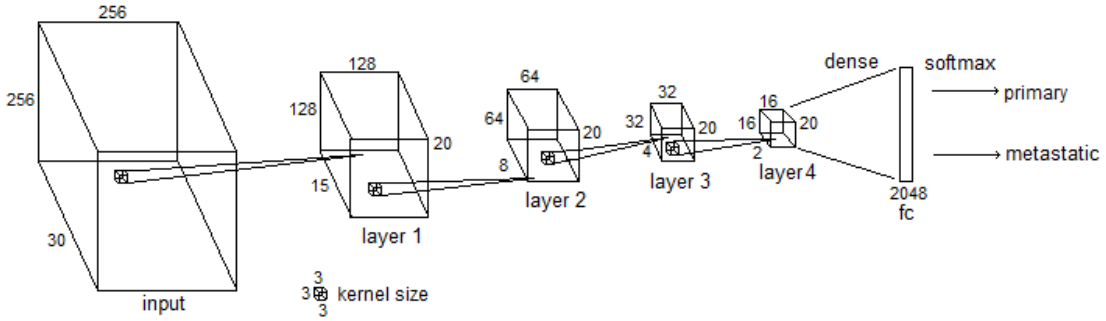

Figure 43 - Training the 3D CNN

Metrics like training loss and accuracy were juxtaposed to their validation counterparts. It is worth mentioning that the number of strided convolutions reflects the need to preserve, to an extent, the cancer regions. For instance, more strided convolutions in the 3D network will cause even less feature maps from slices with regions of interest limiting significantly the ability of converging to a discriminative model.

Most of the other hyperparameters such as the optimizer, the initializer, the size of filters, the activation function or the stride were determined by the literature research following the best practices in the field. Tweaking those has almost no or little impact in this network's behavior or performance.

5-4-4. Support Vector Machines for Evaluating the Extracted Features

Linear SVM is selected as classifier for the new pipeline. SVMs are widely used on cancer discrimination problems [152] since they demonstrate good performance and require fewer data due to its kernel-based nature. The two different approaches, Figure 44, are using the same subset of patients for training and evaluation but with the input data being the differentiating factor. In case of 2D feature extraction each slice is represented by a feature vector contrary to the 3D feature extraction where each feature vector is extracted by the whole volume. The performance metrics include majority vote, for the 2D pipeline, therefore it can be comparable to the performance of the 3D variance.

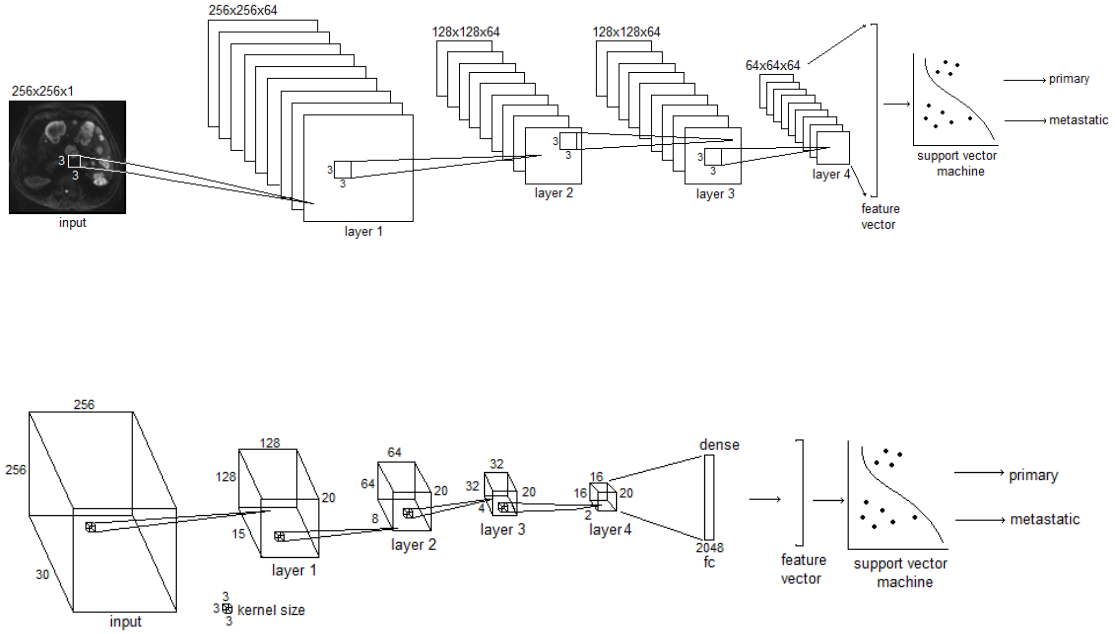

Figure 44 – Extraction and Evaluation of features from 2D(top) and 3D(bottom) networks

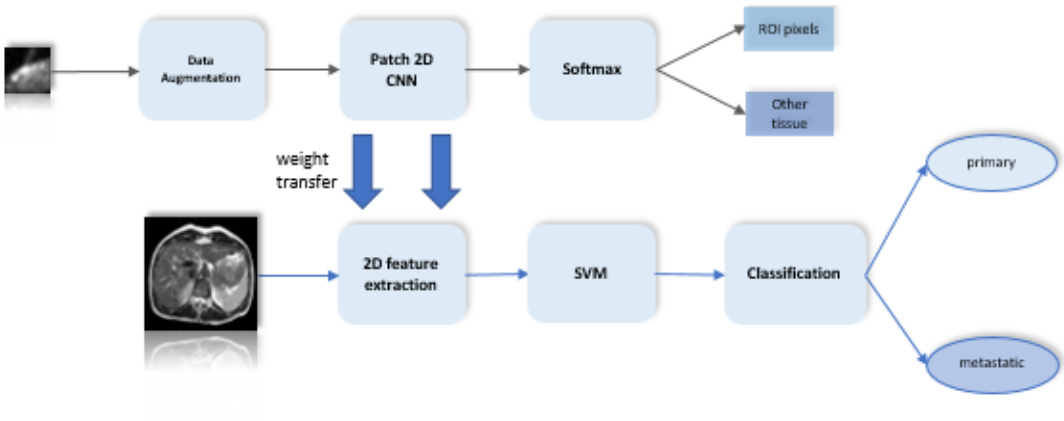

Figure 45 - Training(top) and Feature Extraction(bottom) for the 2D pipeline

5-5. Implementation

As depicted in Figure 45, a patch-wise training strategy is selected to force the network distinguish between malignant and healthy tissue and at the same time exposing the neurons artificially to more data. This is a common practice when there is lack of big datasets or other resources. Since a sufficient model capable of performing well was produced -in this case accuracy 97% and precision 91% in tissue classification- I transfer the weights and the convolutional part of the 2D CNN, which serves as feature extractor for the network, to a new model.

In case of 3D CNN, the feature extractor is trained, and the new model is transferred with the new classifier, bottom pipeline in Figure 46, in a similar way. The key difference is the training strategy where instead of patch-wise training, the full volume is used both in training and inference. This allows to preserve more layers including the fully-connected layer and discarding only the softmax classification layer.

During inference in both approaches the linear SVM is trained in a supervised manner for the new classification problem primary vs metastatic. As mentioned in Section 4-4-4, SVMs are effective with small datasets [102] where only the support vectors are considered for separating the classes by constructing the hyperplane.

5-6. Performance Evaluation

Both 2D and 3D CNN were trained on equivalent subset of the entire cohort, selected from stratified random sampling. A total of 111 patients, 38% primary and 62% metastatic, were included in the initial dataset while 4 were rejected as having less than 30 slices in total, a technical requirement related to the 3D CNN's fixed input extensively discussed in Section 4-3-2. No other clinical criterion for selecting the samples was applied. As a result, a total of 75 patients were used for training and 32 patients for the test set before applying data augmentation, as suggested in Section 4-3-3. In both cases each sample, volume for 3D and slice for 2D feature extractor, was labeled by the medical diagnosis primary or metastatic.

The patch-wise 2D was trained for 30 epochs with approximate training time 45 minutes to 1 hour on a Cuda-compatible Graphics Process Unit (GPU). The model was able to discriminate among malignant and healthy tissue patches with accuracy of 96,9% and precision of 91%.

For the 2D feature extractor: I transfer from the patch-trained network only weights related to the convolutional layers, append the new input 256x256 and discard the fully-connected and softmax layers, Figure 45.The 3D network was trained on volumetric DW-MRI

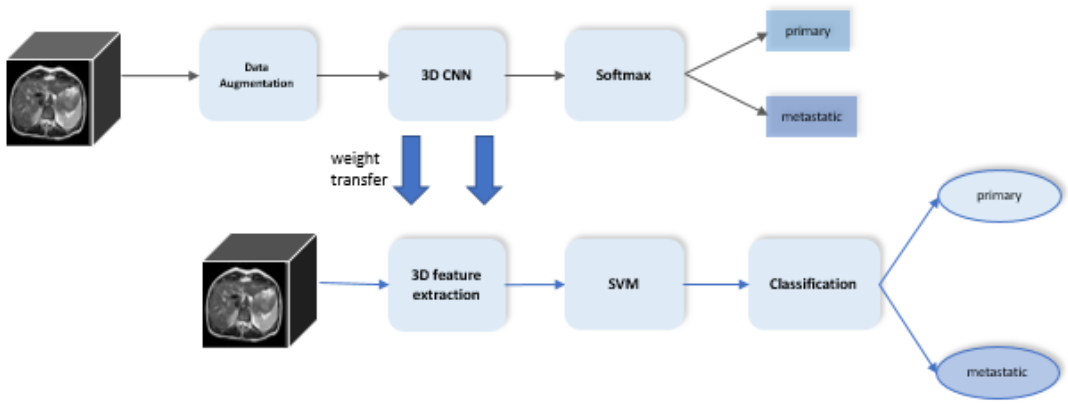

Figure 46 - Training(top) and Feature Extraction(bottom) for the 3D pipeline

	Features - 3D network	Features - 2D network
Accuracy	93.9%	60% (Majority Vote: 65%)
Average precision-recall	98.9%	80%
AUROC	93.7%	53%
Specificity	92.3%	39%
Sensitivity	95%	67%
Precision	94%	66%
Recall - Sensitivity	94%	60%
f1-score	94%	62%

Table 3 - Metrics from the SVM classification comparing features extracted from 3D and 2D networks

data for a total of 20 epochs using the CPU, due to memory limitations on the GPU side, requiring more than 40 hours for training. Similarly to the 2D feature extractor, I transfer the layers and weights including the input up to the fully-connected layer, discarding only the softmax layer, Figure 46.

A linear SVM was trained with the extracted feature vectors from both architectures, from the whole volume in case of the proposed 3D network and from 2D slices of the examined 2D network. The exact subsets of patients used previously for training and testing the CNNs were further equipped the SVM with features for training and testing and the performance of the classifier as presented in Table 3.

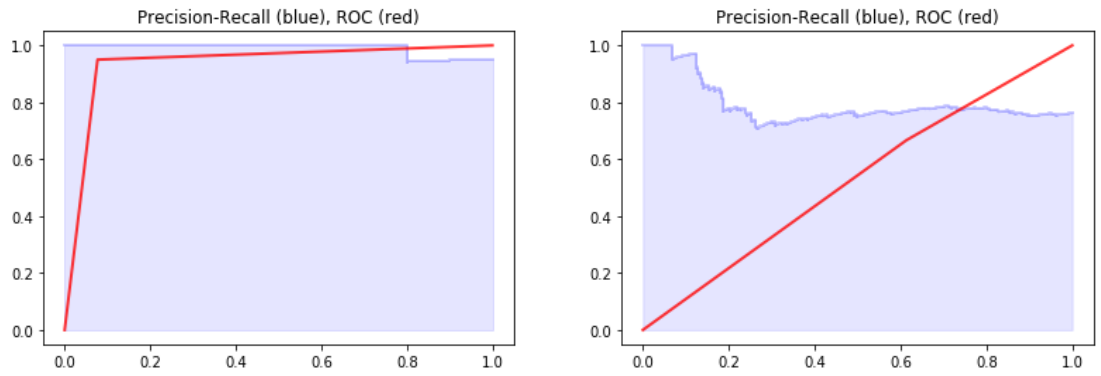

Figure 47 - ROC and Precision-Recall curves for features of 3D(left) and 2D(right) CNN

For binary classification metrics like accuracy could possibly mislead to false conclusions about the performance of each implementation. Thus, the evaluation of performance was assessed by the AUROC and several other metrics, illustrated at Table 3, including not only accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ but also sensitivity-recall $\frac{TP}{TP+FN}$, specificity $\frac{TN}{TN+FP}$, precision $\frac{TP}{TP+FP}$ and F1-score $\frac{2TP}{2TP+FP+FN}$. The notation TP, TN, FP, and FN stand for True Positive, True Negative, False Positive and False Negative respectively.

These predictions were retrieved from the confusion matrices for each pipeline, Figure 48. Additionally, qualitative representation of the performance was given by the ROC and the precision-recall curve in Figure 47. For medical related problems it is worth noting that sensitivity is important since it make sure the actual positive (metastatic) samples are accurately classified.

The calculated confusion matrix at the testing phase is given by Figure 48. The combined 3D feature extractor - SVM pipeline classified correctly 11 out of 12 patients from the primary tumor class and 19 out of 20 belonging to the metastatic class. It is worth noting that the networks have different types of inputs, volume for the 3D and individual slices for the 2D feature extractor, therefore by extension the two matrices are not directly comparable between them.

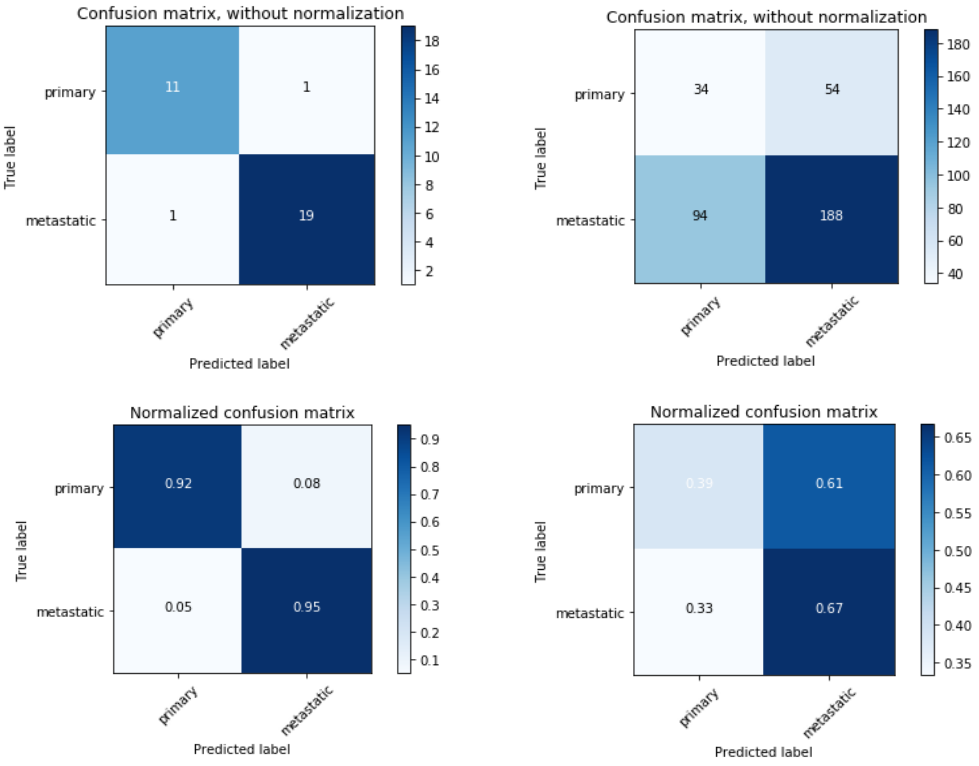

Figure 48 - Confusion matrices 3D(left) and 2D(right) pipeline

6. Conclusion

In recent years there has been a rising interest in research related to medical image applications that leverage DL techniques for various clinical conditions, following the successful paradigm of natural images in real life applications. With regards to cancer research broad horizons are emerging by the exploitation of novel architectures, engineering new feature representations and by taking advantage the special traits and properties of the domain.

The proposed automated tissue characterization by the extracted CNN features are evaluated in different network architectures, training strategies and using data augmentation to inflate the number of samples and simultaneously deal with class imbalance issues related to the dataset. The tomographic DW-MRI data provide enhanced anatomical information about the region of interest. The consecutive convolutional layers produced feature maps based on the interactions of learnt kernels and the input slices in order to amplify the most discriminative areas and capture the nonlinear tenor of each class.

Additionally, the proposed 3D CNN takes advantage of the intra and inter-slice tumor relationships resulting in a richer representation which gradually captures the inner and overall structure of each tumor type. The performance evaluation shows that voxel-wise architectures outperform the 2D variants by a wide margin, with a 93.4% to 65% accuracy and sensitivity of 95% to 67%, in liver tumor tissue classification (primary vs. metastatic) using the same dataset.

This significant improvement is a result of multiple factors including: a)the image representation inside the 3D network is far superior and this is indicative from the larger amount of learnable parameters, b) the relation of the acquired knowledge from the model and the dataset is more complicated mainly due to voxel kernels, c) the inclusion of the dense layer in the feature extraction architecture provides feature selection properties in an “organic” way, d) volumetric features from whole slices are more suitable for this clinical application due to the high variance in tumor structure spanning from compact regions to dispersed small tumors across large areas, including metastases for many different organs.

The aim of this research was to integrate the extracted features into a generic framework with hand-crafted elements assisting in difficult problems, like in this case study, the intraclass discrimination among liver cancer tissue. To the best of our knowledge this is the first time where 3D strided CNN was successfully applied for cancer feature extraction from the whole volume with no need of segmenting or detecting the regions of interest first.

Bibliography

[1] E. Hosseini-Asl, J. M. Zurada and O. Nasraoui, "Deep Learning of Part-Based Representation of Data Using Sparse Autoencoders With Nonnegativity Constraints," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 27, no. 12, pp. 2486 - 2498, 2016.

[2] A. Lemme, R. F. Reinhart and J. J. Steil, "Online learning and generalization of parts-based image representations by non-negative sparse autoencoders," *Neural Networks,* vol. 33, pp. 194-203, 2012.

[3] A. Ng, "Sparse autoencoder," [Online]. Available: https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf. [Accessed 01 05 2018].

[4] H. Lee, R. Grosse, R. Ranganath and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, Montreal, 2009.

[5] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," 2010. [Online]. Available: http://www.cs.toronto.edu/%7Ekriz/conv-cifar10-aug2010.pdf. [Accessed 02 02 2018].

[6] M. Egmont-Petersen, D. de Ridder and H. Handels, "Image processing with neural networks - a review," *Pattern Recognition,* vol. 35, no. 10, pp. 2279-2301, 2002.

[7] F. Visin, M. Ciccone and A. Romero, "ReSeg: A Recurrent Neural Network-based Model for Semantic Segmentation," in *Computer Vision and Pattern Recognition*, Las Vegas, 2016.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[9] F. Visin, K. Kastner and K. Cho, "ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks," *arXiv preprint arXiv:1505.00393,* 2015.

[10] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Neural Information Processing Systems*, Vancouver, 2009.

[11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation,* vol. 1, no. 4, pp. 541-551, 1989.

[12] Y. LeCun, F. J. Huang and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," *Proceedings of the Computer Vision and Pattern Recognition (CVPR),* pp. 96-104, 2004.

[13] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin and S. Yan, "Human parsing with contextualized convolutional neural network," *Proceedings of the IEEE International Conference on Computer Vision,* pp. 1386-1394, 2015.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,* pp. 1725-1732, 2014.

[15] K. Jarrett, K. Kavukcuoglu and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *IEEE 12th International Conference on Computer Vision*, Pittsburgh, 2009.

[16] Y. Xiang, W. Choi and Y. Lin, "Subcategory-aware Convolutional Neural Networks for Object Proposals and Detection," in *Winter Conference on Applications of Computer Vision*, Santa Rosa, 2017.

[17] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," *International Conference on Intelligent Robots and Systems (IROS),* pp. 922-928, 2015.

[18] S. Gupta, R. Girshick, P. Arbelaez and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *European Conference on Computer Vision*, Cham, Springer, 2014, pp. 345-360.

[19] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Computer Vision and Pattern Recognition*, Boston, 2015.

[20] V. Badrinarayanan, A. Kendal and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *arXiv preprint arXiv:1511.00561,* 2016.

[21] B. Hariharan, P. Arbelaez, R. Girshick and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*, Cham, Springer, 2014, pp. 297-312.

[22] I. Goodfellow, J. Pouget-Abadie and M. Mirza, "Generative Adversarial Nets," in *Neural Information Processing Systems*, Montreal, 2014.

[23] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 8 April 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. [Accessed 8 April 2018].

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Cham, Springer, 2014, pp. 818-833.

[25] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Fisher Networks for Large-Scale Image Classification," in *Neural Information Processing Systems*, Lake Tahoe, 2013.

[26] K. Simonyan, O. M. Parkhi, A. Vedaldi and A. Zisserman, "Fisher Vector Faces in the Wild," in *British Machine Vision Conference*, Bristol, 2013.

[27] J. Krapac, J. Verbeek and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *International Conference on Computer Vision*, Barcelona, 2011.

[28] J.-G. Wang, J. Li and W.-Y. Yau, "Boosting Dence SIFT Descriptors and Shape Contexts of face images for gender recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Rhode Island, 2010.

[29] T. Vo, D. Tran and W. Ma, "Tensor Decomposition of Dense SIFT Descriptors in Object Recognition," in *European Symposium on Artificial Neural Networks, Computational Intelligence*, Bruges, 2014.

[30] F. Perronnin, J. Sánchez and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *European Conference on Computer Vision*, Heraklion, 2010.

[31] M. Tanaka, A. Torii and M. Okutomi, "Fisher Vector based on Full-covariance Gaussian Mixture Model," *Information and Media Technologies,* vol. 8, no. 4, pp. 1041-1045, 2013.

[32] G. Csurka and F. Perronnin, "Fisher Vectors: Beyond Bag-of-Visual-Words Image Representations," in *International Conference on Computer Vision, Imaging and Computer Graphics*, Angers, 2010.

[33] V. Sydorov, M. Sakurada and C. H. Lampert, "Deep Fisher Kernels - End to End Learning of the Fisher Kernel GMM Parameters," in *Computer Vision and Pattern Recognition*, Columbus, 2014.

[34] K. Chatfield, K. Simonyan and A. Vedaldi, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," *arXiv preprint arXiv:1405.3531,* 2014.

[35] L. Zheng, Y. Yang and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval," *Pattern Analysis and Machine Intelligence,* vol. PP, no. 99, p. 1, 2017.

[36] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556,* 2015.

[37] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki and S. Carlsson, "From Generic to Specific Deep Representations for Visual Recognition," in *CVPRW DeepVision Workshop*, Boston, 2015.

[38] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems*, Lake Tahoe, 2012.

[39] C. Szegedy, W. Liu and Y. Jia, "Going Deeper with Convolutions," in *Conference on Computer Vision and Pattern Recognition*, Boston, 2015.

[40] R. K. Srivastava, K. Greff and J. Schmidhuber, "Training Very Deep Networks," *Advances in neural information processing systems,* pp. 2377-2385, 2015.

[41] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Computer Vision and Pattern Recognition*, Las Vegas, 2016.

[42] S. Targ, D. Almeida and K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," *arXiv preprint arXiv:1603.08029,* 2016.

[43] R. K. Srivastava, K. Greff and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387,* 2015.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, 2017.

[45] P. Pinheiro, R. Collobert and P. Dollar, "Learning to segment object candidates," *Advances in Neural Information Processing Systems,* pp. 1990-1998, 2015.

[46] H. Noh, S. Hong and B. Han, "Learning deconvolution network for semantic segmentation," *Proceedings of the IEEE International Conference on Computer Vision,* pp. 1520-1528, 2015.

[47] J. Dai, K. He and Y. Li, "Instance-sensitive Fully Convolutional Networks," in *European Conference on Computer Vision*, Amsterdam, 2016.

[48] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected," in *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2018, pp. 834-848.

[49] S. Hong, H. Noh and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation.," in *Advances in neural information processing systems*, Montreal, 2015.

[50] D. Kiela and L. Bottou, "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics," in *Empirical Methods on Natural Language Processing*, Doha, 2014.

[51] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781 ,* 2013.

[52] L. Ma, Z. Lu, L. Shang and H. Li, "Multimodal Convolutional Neural Networks for Matching Image and Sentence," in *International Conference on Computer Vision*, Santiago, 2015.

[53] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Neural Information Processing Systems*, Montreal, 2014.

[54] X. Zhang, H. Xiong, W. Zhou, W. Lin and Q. Tian, "Picking Deep Filter Responses for Fine-grained Image Recognition," in *Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016.

[55] G. Amato, F. Falchi, F. Rabitti and L. Vadicamo, "Combining Fisher Vector and Convolutional Neural Networks for image retrieval," in *CEUR Workshop Proccedings*, Atlanta, 2016.

[56] M. Cimpoi, S. Maji and A. Vedaldi, "Deep Filter Banks for Texture Recognition and Segmentation," in *Computer Vision and Pattern Recognition*, Boston, 2015.

[57] D. Erhan, C. Szegedy, A. Toshev and D. Anguelov, "Scalable Object Detection using Deep Neural Networks," in *Computer Vision and Pattern Recognition*, Columbus, 2014.

[58] C. Szegedy, S. Reed and D. Erhan, "Scalable High Quality Object Detection," *arXiv preprint arXiv:1412.1441,* 2015.

[59] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The KITTI dataset.," *The International Journal of Robotics Research,* vol. 32, no. 11, pp. 1231-1237, 2013.

[60] G. Brostow, J. Fauqueur and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters,* vol. 30, no. 2, pp. 88-97, 2009.

[61] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with

region proposal networks," in *Advances in neural information processing systems*, Montreal, 2015.

[62] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 580-587, 2014.

[63] Y. Gong, L. Wang, R. Guo and S. Lazebnik, "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," in *European Conference on Computer Vision*, Zurich, 2014.

[64] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, "Overfeat: Integrated recognition, localization and detection," *arXiv preprint arXiv:1312.6229,* 2013.

[65] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN Features off-the-shelf: An Astounding Baseline for Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Columbus, 2014.

[66] O. Penatti, K. Nogueira and J. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, 2015.

[67] V. Andrearczyk and P. F. Whelan, "Using Filter Banks in Convolutional Neural Networks for Texture Classification," *Pattern Recognition Letters,* vol. 84, pp. 63-69, 1 December 2016.

[68] C. Bailer, T. Habtegebrial, K. Varanasi and D. Stricker, "Fast Dense Feature Extraction with CNNs that have Pooling or Striding Layers," [Online]. Available: https://www.researchgate.net/profile/Christian_Bailer/publication/319551162_Fast_Dense_Feature_Extraction_with_CNNs_that_have_Pooling_or_Striding_Layers/links/59b3063d0f7e9b37434eadbd/Fast-Dense-Feature-Extraction-with-CNNs-that-have-Pooling-or-Striding-L. [Accessed 21 April 2018].

[69] M. Zeiler, G. Taylor and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, 2011.

[70] C. Dong, C. C. Loy, K. He and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, Cham, Springer, 2014, pp. 184-199.

[71] Z. Shi, M. Shi and C. Li, "The prediction of character based on recurrent neural network language model," in *16th International Conference on Computer and Information Science (ICIS)*, Wuhan, 2017.

[72] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078,* 2014.

[73] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[74] A. Oord, N. Kalchbrenner and K. Kavukcuoglu, "Pixel Recurrent Neural Networks," *arXiv preprint arXiv:1601.06759,* 2016.

[75] G. Hinton and R. Zemel, "Autoencoders, minimum description length, and helmholtz free energy," in *Advances in neural information processing systems*, 1994, pp. 3-10.

[76] G. E. Hinton and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science,* vol. 313, no. 5768, pp. 504-507, 2006.

[77] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," *Advances in neural information processing systems,* pp. 809-817, 2013.

[78] J. Masci, U. Meier, D. Cirecsan and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks.*, Berlin, Springer, 2011, pp. 52-59.

[79] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114,* 20 12 2013.

[80] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Iberoamerican Congress on Pattern Recognition*, Berlin, Springer, 2012, pp. 14-36.

[81] G. E. Hinton, S. Osindero and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation,* vol. 18, no. 7, pp. 1527-1554, 2006.

[82] A. Antoniou, A. Storkey and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv preprint arXiv:1711.04340,* 2017.

[83] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang and D. Metaxas, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," *arXiv preprint arXiv:1710.10916,* 2017.

[84] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, "Generative Adversarial Text to Image Synthesis," *arXiv preprint arXiv:1605.05396,* 2016.

[85] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, "druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico," *Molecular Pharmaceutics,* vol. 14, no. 9, p. 3098–3104, 2017.

[86] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research,* vol. 11, no. Feb, pp. 625-660, 2010.

[87] R. Raina, A. Battle, H. Lee, B. Packer and A. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*, ACM, 2007, pp. 759-766.

[88] A. Saxe, J. McClelland and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120,* 2013.

[89] Y. Bengio, P. Lamblin and D. Popovici, "Greedy Layer-Wise Training of Deep Networks," in *Neural Information Processing Systems*, Vancouver, 2007.

[90] N. Tajbakhsh, J. Y. Shin and S. R. Gurudu, "Convolutional Neural Networks for Medical Image

Analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1299 - 1312, 2016.

[91] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv preprint arXiv:1412.6572,* 2014.

[92] A. Nguyen, J. Yosinski and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015.

[93] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International journal of computer vision,* vol. 111, no. 1, pp. 98-136, 2015.

[94] O. Russakovsky, J. Deng, H. Su, J. Krause and S. Satheesh, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.

[95] B. Russell, A. Torralba, K. Murphy and W. Freeman, "LabelMe: a database and web-based tool for image annotation," *International journal of computer vision,* vol. 77, no. 1-3, pp. 157-173, 2008.

[96] S. Song, S. Lichtenberg and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *CVPR*, Boston, 2015.

[97] Y. LeCun, C. Cortes and C. J. Burges, "The MNIST Dataset Of Handwritten Digits," [Online]. Available: http://yann.lecun.com/exdb/mnist/. [Accessed 1 May 2018].

[98] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 248-255, 2009.

[99] B. H. Menze, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging,* vol. 34, no. 10, pp. 1993-2024, 2015.

[100] H.-C. Shin, H. R. Roth, M. Gao, L. Lu and Z. Xu, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *Medical Imaging,* vol. 5, no. 35, pp. 1285 - 1298, 2016.

[101] J. Xu, X. Luo, G. Wang, H. Gilmore and A. Madabhushi, "A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing,* vol. 191, pp. 214-223, 2016.

[102] G. Amit, R. Ben-Ari, O. Hadad, E. Monovich, N. Granot and S. Hashoul, "Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches," in *Medical Imaging*, Orlando, 2017.

[103] Q. Li, W. Cai, X. Wang and Y. Zhou, "Medical Image Classification with Convolutional Neural Network," in *Automation, Robotics & Vision*, Singapore, 2014.

[104] M. Gao, U. Bagci, L. Lu, A. Wu and G. Papadakis, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in*

Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 6, no. 1, pp. 1-6, 2018.

[105] W. Shen, M. Zhou, F. Yang, C. Yang and J. Tian, "Multi-scale Convolutional Neural Networks for Lung Nodule Classification," in *Information Processing in Medical Imaging*, Isle of Skye, 2015.

[106] F. Ciompi, B. de Hoop, S. van Riel, K. Chung and E. Scholten, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Medical image analysis,* vol. 26, no. 1, pp. 195-202, 2015.

[107] J. Wolterink, T. Leiner and M. Viergever, "Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 2015.

[108] D. Yang, T. Xiong, D. Xu and Q. Huang, "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization," in *International Conference on Information Processing in Medical Imaging*, Cham, Springer, 2017, pp. 633-644.

[109] E. Hosseini-Asl, R. Keynton and A. El-Baz, "Alzheimer's disease diagnostics by adaptation of 3D convolutional network," in *Image Processing*, Phoenix, 2016.

[110] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks," *arXiv preprint arXiv:1502.02506,* 2015.

[111] V. Gulshan, L. Peng and M. Coram, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *Jama,* vol. 316, no. 22, pp. 2402-2410, 13 December 2016.

[112] B. D. de Vos, J. M. Wolterink, P. A. de Jong, M. A. Viergever and I. Išgum, "2D image classification for 3D anatomy localization: employing deep convolutional neural networks," in *Medical Imaging 2016: Image Processing*, San Diego, 2016.

[113] S. Kohl, D. Bonekamp and H.-P. Schlemmer, "Adversarial Networks for the Detection of Aggressive Prostate Cancer," *arXiv preprint arXiv:1702.08014,* 26 February 2017.

[114] H. Roth, L. Lu, A. Seff, K. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey and R. Summers, "A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2014, pp. 520-527.

[115] H.-C. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35, no. 5, pp. 1285-1298, 2016.

[116] . A. A. A. Setio, F. Ciompi and G. Litjens, "Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1160 - 1169, 2016.

[117] Q. Dou, H. Chen, L. Yu, J. Qin and P.-A. Heng, "Multi-level Contextual 3D CNNs for False Positive Reduction in Pulmonary Nodule Detection," *IEEE Transactions on Biomedical Engineering,* vol. 64,

no. 7, pp. 1558-1567, 2017.

[118] G. Maicas, G. Carneiro and A. Bradley, "Globally optimal breast mass segmentation from DCE-MRI using deep semantic segmentation as shape prior," in *International Symposium on Biomedical Imaging*, Melbourne, 2017.

[119] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 20-28, 2015.

[120] I. Ivsgum, M. Benders, B. Avants and M. Cardoso, "Evaluation of automatic neonatal brain segmentation algorithms," *Medical image analysis,* vol. 20, no. 1, pp. 135-151, 2015.

[121] H. Roth, L. Lu and A. Farag, "DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, Munich, 2015.

[122] J. Cai, L. Lu and Y. Xie, "Pancreas Segmentation in MRI Using Graph-Based Decision Fusion on Convolutional Neural Networks," in *Medical Image Computing and Computer-Assisted*, Quebec, 2017.

[123] H. Roth, A. Farag, L. Lu, E. Turkbey and R. Summers, "Deep convolutional networks for pancreas segmentation in CT imaging," in *Medical Imaging 2015: Image Processing*, vol. 9413, International Society for Optics and Photonics, 2015, p. 94131G.

[124] L. Yu, X. Yang, H. Chen, J. Qin and P.-A. Heng, "Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images," in *AAAI Conference on Artificial Intelligence*, San Francisco, 2017.

[125] P. V. Tran, "A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI," *arXiv preprint arXiv:1604.00494,* 27 April 2016.

[126] R. Poudel, P. Lamata and G. Montana, "Recurrent Fully Convolutional Neural Networks for Multi-slice MRI Cardiac Segmentation," in *International Workshop on Reconstruction and Analysis of Moving Body Organs*, Athens, 2016.

[127] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam and M. Nielsen, "Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network," in *Medical Image Computing and Computer-Assisted*, Nagoya, 2013.

[128] S. Turaga, J. Murray, V. Jain and F. Roth, "Convolutional networks can learn to generate affinity graphs for image segmentation," *Neural computation,* vol. 22, no. 2, pp. 511-538, 2010.

[129] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara and H. Fujita, "First trial and evaluation of anatomical structure segmentations in 3D CT images based only on deep learning," *Medical Imaging and Information Sciences,* vol. 33, no. 3, pp. 69-74, 2016.

[130] W. Crum, O. Camara and D. Hill, "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis," *IEEE transactions on medical imaging,* vol. 25, no. 11, pp. 1451-1461, 2006.

[131] K. Kamnitsas, C. Ledig and V. F. Newcombe, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis,* vol. 36, pp. 61-78, 2017.

[132] W. Zhang, R. Li and H. Deng, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage,* pp. 214-224, March 2015.

[133] P. Moeskops, M. A. Viergever and A. M. Mendrik, "Automatic segmentation of MR brain images with a convolutional neural network," *Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1252 - 1261, 2016.

[134] H. Dong, G. Yang, F. Liu, Y. Mo and Y. Guo, "Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks," in *Medical Image Understanding and Analysis*, Edinburgh, 2017.

[135] S. Pereira, A. Pinto, V. Alves and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1240 - 1251, 2016.

[136] M. Havaei, A. Davy and D. Warde-Farley, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis,* vol. 35, pp. 18-31, 2017.

[137] M. Drozdzal, E. Vorontsov and G. Chartrand, "The Importance of Skip Connections in Image Segmentation," in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, Athens, 2016.

[138] H. Chen, X. Qi and J.-Z. Cheng, "Deep Contextual Networks for Neuronal Structure Segmentation," in *AAAI Conference on Artificial Intelligence*, Phoenix, 2016.

[139] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer*, Munich, 2015.

[140] Ö. Çiçek, A. Abdulkadir and S. S. Lienkamp, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted*, Istanbul, 2016.

[141] F. Milletari, N. Navab and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Fourth International Conference on 3D Vision*, Stanford, 2016.

[142] H. Chen, Q. Dou and L. Yu, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage,* 18 April 2017.

[143] P. F. Christ, M. Elshaer and F. Ettlinger, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Medical Image Computing and Computer-Assisted Intervention*, Istanbul, 2016.

[144] J. Chen, L. Yang and Y. Zhang, "Combining Fully Convolutional and Recurrent Neural Networks for 3D Biomedical Image Segmentation," in *Neural Information Processing Systems*, 2016.

[145] M. Stollenga, W. Byeon and M. Liwicki, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," in *Neural Information Processing Systems*,

Barcelona , 2015.

[146] M. Avetisian, "Volumetric Medical Image Segmentation with Deep Convolutional Neural Networks," [Online]. Available: http://ceur-ws.org/Vol-2022/paper03.pdf. [Accessed 5 1 2018].

[147] L. Wu, X. Yang and S. Li, "Cascaded Fully Convolutional Networks for automatic prenatal ultrasound image segmentation," in *International Symposium on Biomedical Imaging*, Melbourne, 2017.

[148] M. Ghafoorian, A. Mehrtash and T. Kapur, "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Cham, Springer, 2017, pp. 516-524.

[149] H. Chen, X. Qi, L. Yu and P.-A. Heng, "DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation," in *Computer Vision and Pattern Recognition*, Las Vegas, 2016.

[150] Y. Bar, I. Diamant, L. Wolf and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," in *Medical Imaging 2015: Computer-Aided Diagnosis*, International Society for Optics and Photonics, 2015, p. 94140V.

[151] Z. Li, Y. Wang, J. Yu, Y. Guo and W. Cao, "Deep Learning based Radiomics," *Nature,* vol. 7, no. Scientific Reports, 2017.

[152] T. S. Furey, N. Cristianini and N. Duffy, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics,* vol. 16, no. 10, p. 906–914, 2000.

[153] A. Namburete, W. Xie, M. Yaqub, A. Zisserman and A. Noble, "Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning," *Medical image analysis,* vol. 46, pp. 1-14, 2018.

[154] R. J. Gillies, P. E. Kinahan and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology,* vol. 278, no. 2, pp. 563-577, 2016.

[155] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker and H. J. Aerts, "Radiomics: extracting more information from medical images using advanced feature analysis.," *European Journal of Cancer,* vol. 48, no. 4, pp. 441-446, 2012.

[156] M. Vallières, C. R. Freeman, S. R. Skamene and I. E. Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine & Biology,* vol. 60, no. 14, p. 5471–5496, 2015.

[157] W. Aichinger, S. Krappe, A. E. Cetin, R. Cetin-Atalay, A. Uner, M. Benz, T. Wittenberg, M. Stamminger and C. Munzenmayer, "Automated cancer stem cell recognition in H and E stained tissue using convolutional neural networks and color deconvolution," *Medical Imaging 2017: Digital Pathology,* vol. 10140, p. 101400N, 2017.

[158] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "Modeling the Intra-class Variability for Liver Lesion Detection Using a Multi-class Patch-Based CNN,"

International Workshop on Patch-based Techniques in Medical Imaging, pp. 129-137, 2017.

[159] K. Yasaka, H. Akai, O. Abe and S. Kiryu, "Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study," *Radiology,* p. 170706, 2017.

[160] L. Bi, J. Kim, A. Kumar and D. Feng, "Automatic Liver Lesion Detection using Cascaded Deep Residual Networks," *arXiv preprint arXiv:1704.02703,* 2017.

[161] R. Bharath and R. Rajalakshmi, "Deep scattering convolution network based features for ultrasonic fatty liver tissue characterization," in *39th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC)*, Jeju Island, 2017.

[162] A. Jorgensen, J. Fagertun and T. Moeslund, "Diagnosis of Broiler Livers by Classifying Image Patches," *Scandinavian Conference on Image Analysis,* pp. 374-385, 2017.

[163] X. Liu, J. L. Song, S. H. Wang, J. W. Zhao and Y. Q. Chen, "Learning to Diagnose Cirrhosis with Liver Capsule Guided Ultrasound Image Classification," *Sensors,* vol. 17, no. 1, pp. 149-160, 2017.

[164] E. Vorontsov, A. Tang, D. Roy, C. J. Pal and S. Kadoury, "Metastatic liver tumour segmentation with a neural network-guided 3D deformable model," *Medical & biological engineering & computing,* vol. 55, no. 1, pp. 127-139, 2017.

[165] E. Kang, J. Min and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction," *Medical physics,* vol. 44, no. 10, 2017.

[166] P. F. Christ, F. Ettlinger and G. Kaissis, "SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D Convolutional Neural Networks," *IEEE 14th International Symposium on Biomedical Imaging (ISBI),* pp. 839-843, 2017.

[167] A. Ben-Cohen, E. Klang, I. Diamant, N. Rozendorn, S. P. Raskin, E. Konen, M. M. Amitai and H. Greenspan, "CT Image-based Decision Support System for Categorization of Liver Metastases Into Primary Cancer Sites: Initial Results," *Academic radiology,* vol. 24, no. 12, pp. 1501-1509, 2017.

[168] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas and D. Comaniciu, "Automatic Liver Segmentation Using an Adversarial Image-to-Image Network," *International Conference on Medical Image Computing and Computer-Assisted Intervention,* pp. 507-515, 2017.

[169] W. Li, F. Jia and Q. Hu, "Automatic segmentation of liver tumor in CT images with deep convolutional neural networks," *Journal of Computer and Communications,* vol. 3, no. 11, pp. 146-151, 2015.

[170] C.-L. Kuo, S.-C. Cheng, C.-L. Lin, K.-F. Hsiao and S.-H. Lee, "Texture-based treatment prediction by automatic liver tumor segmentation on computed tomography," *International Conference on Computer Information and Telecommunication Systems (CITS),* pp. 128-132, 2017.

[171] P. F. Christ, M. E. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann and M. D'Anastasi, "Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields," *International Conference on Medical Image Computing and Computer-Assisted Intervention,* pp.

415-423, 2016.

[172] P. F. Christ, F. Ettlinger, F. Grun, M. E. Elshaera, J. Lipkova and S. Schlecht, "Automatic Liver and Tumor Segmentation of CT and MRI Volumes using Cascaded Fully Convolutional Neural Networks," *arXiv preprint arXiv:1702.05970,* 2017.

[173] C. Sun, S. Guo, H. Zhang, J. Li, M. Chen, S. Ma, L. Jin, X. Liu, X. Li and X. Qian, "Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs," *Artificial intelligence in medicine,* vol. 83, pp. 58-66, 2017.

[174] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang and B. Hu, "Liver fibrosis classification based on transfer learning and FCNet for ultrasound images," *IEEE Access,* vol. 5, pp. 5804-5810, 2017.

[175] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai and H. Greenspan, "Fully Convolutional Network for Liver Segmentation and Lesions Detection," *Deep Learning and Data Labeling for Medical Applications,* pp. 77-85, 2016.

[176] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu and P. A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Liver Tumor Segmentation from CT Volumes," *arXiv preprint arXiv:1709.07330,* 2017.

[177] E. Vorontsov, G. Chartrand, A. Tang, C. Pal and S. Kadoury, "Liver lesion segmentation informed by joint liver segmentation," *arXiv preprint arXiv:1707.07734v2,* 2017.

[178] Y. Yuan, "Hierarchical Convolutional-Deconvolutional Neural Networks for Automatic Liver and Tumor Segmentation," *arXiv preprint arXiv:1710.04540,* 2017.

[179] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin and P.-A. Heng, "3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes," *International Conference on Medical Image Computing and Computer-Assisted Intervention,* pp. 149-157, 2016.

[180] A. Hoogi, J. W. Lambert, Y. Zheng, D. Comaniciu and D. L. Rubin, "A Fully-Automated Pipeline for Detection and Segmentation of Liver Lesions and Pathological Lymph Nodes," *arXiv preprint arXiv:1703.06418,* 2017.

[181] X. Han, "Automatic Liver Lesion Segmentation Using A Deep Convolutional Neural Network Method," *arXiv preprint arXiv:1703.06418,* 2017.

[182] L. Chen, Y. Wu, A. DSouza, A. Abidin, A. Wismuller and C. Xu, "MRI Tumor Segmentation with Densely Connected 3D CNN," *Medical Imaging 2018: Image Processing,* vol. 10574, p. 105741F, 2018.

[183] P. Hu, F. Wu, J. Peng, P. Liang and D. Kong, "Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution," *Physics in Medicine & Biology,* vol. 61, no. 24, p. 8676, 2016.

[184] F. Lu, F. Wu, P. Hu, Z. Peng and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *International journal of computer assisted radiology and surgery,* vol. 12, no. 2, pp. 171-182, 2017.

[185] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167,* 2015.

[186] B. Gaonkar, D. Hovda, N. Martin and L. Macyszyn, "Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation," *Medical Imaging 2016: Computer-Aided Diagnosis,* vol. 9785, p. 97852I, 2016.

[187] M. Abadi and e. al., ""TensorFlow: A System for Large-Scale Machine Learning."," in *13th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, 2016.

[188] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. W. Fisher III and L. K. Hansen, "Dreaming More Data: Class-dependent Distributions over Diffeomorphisms for Learned Data Augmentation," *Artificial Intelligence and Statistics,* pp. 342-350, May 2016.

[189] X.-S. Wei, "Must Know Tips/Tricks in Deep Neural Networks," 19 October 2015. [Online]. Available: http://lamda.nju.edu.cn/weixs/project/CNNTricks/CNNTricks.html. [Accessed 12 May 2018].

[190] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv preprint arXiv:1712.04621,* 13 December 2017.

[191] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, Washington, DC, 2015.

[192] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806 ,* 2014.

[193] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580,* 2012.

[194] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the thirteenth international conference on artificial intelligence and statistics,* pp. 249-256, 2010.

[195] K. P. Diederik and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980v9,* 2014.

[196] Y. D. Cid, J. Castelli, R. Schaer, N. Scher, A. Pomoni, J. Prior and A. Depeursinge, "QuantImage: An Online Tool for High-Throughput 3D Radiomics Feature Extraction in PET-CT," *Biomedical Texture Analysis,* pp. 349-377, 2018.

[197] M. Chang, 21 November 2016. [Online]. Available: https://www.slideshare.net/ckmarkohchang/generative-adversarial-networks. [Accessed 15 July 2018].

Index