Literature Review

XCS229ii - Machine Learning Strategy and Intro to Reinforcement Learning

Stanford Center For Professional Development

By Akshay Agarwal, Manish Das, Jaro Habr, Parag Kanade

1 General Problem and Task Definition

Image segmentation is the process of partitioning an image into multiple segments (sets of pixels, also known as image objects). In the medical domain, Image segmentation is considered the most essential process as it extracts the region of interest (ROI) through a semi automatic or automatic process. It divides an image into areas based on a specified description, such as segmenting body organs/tissues in the medical applications for border detection, tumor detection/segmentation, and mass detection. Medical Image segmentation has a lot of applications such as locating tumors, surgery planning, diagnosis, measuring tissue volumes etc (Karimi et al., 2021; Valanarasu et al., 2021).

A lot of the research has been done in image segmentation in the last 20 years. Prior to neural networks, unsupervised ML approaches like k-means clustering were used to segment images. Post 2015, a lot of work has been done using Convolutional Neural Networks (CNNs) to address the problems faced before. Recently Transformer based architectures are also used in image segmentation tasks which has shown encouraging results by achieving State-of-the-Art performance.

The central problem which most of the papers try to address in this field can be listed as below:

1. Reduced feature resolution while encoding and decoding the images
2. Localized Attention due to CNNs network
3. Sparsity of data available in medical imaging
4. Boundary detection in segmentation

Starting with the U-Net (Ronneberger et al., 2015), which is the baseline model for the majority of the research work, the papers were selected based on relevance to medical segmentation, the presented model's performance as well as our goal to learn about and cover the latest research direction in the aforementioned domain.

2 Article Summaries

In this chapter we summarize twelve papers by briefly describing the main problem they address and by presenting the key ideas and contributions to solve the identified problems. In the first section we cover CNN-based approaches followed by the lately emerged Transformer-based approaches in the second section.

2.1 CNN-based Approach

2.1.1 U-Net: Convolutional Networks for Biomedical Image Segmentation (Ronneberger et al., 2015)

Problem: Fast and efficient biomedical image segmentation using deep learning and very few annotated training images.
Key idea: In medical image segmentation the goal is to assign each pixel in the image a label in order to separate different organs or tissues captured within the image. In this paper, the authors present a new model architecture called the U-Net. U-Net contains a contracting path as well as an expansive path both of which use 3 by 3 convolutions. The image downsampling is achieved by a repetitive application of a 2 by 2 max-pooling layer, which also aids to address the speed issue of the current segmentation models. For image upsampling the authors use 2 by 2 up-convolutions and concatenate them with a corresponding cropped feature map from the contracting path (skip-connections) which helps the model to preserve spatial information and hence localization accuracy. The symmetry of the contracting and expanding paths gives the architecture its u-shaped appearance.
To address the problem where only a few annotated images are available for training the authors apply data augmentation techniques like shift, rotation as well as random elastic deformations to the training samples. This helps the model to learn the desired invariance and robustness properties.

2.1.2 DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs (Chen et al., 2016)

Problem: Main challenges in image segmentation tasks are reduced feature resolution, multiple scales objects present in image and boundary detection.
Key idea: The paper tackles the above challenges by using Upsampled filters (atrous convolution), Atrous Spatial Pyramid Pooling (ASPP) and Conditional Random Field (CRF). Instead of downsampling, it uses upsampling to increase resolution in the last few layers (atrous convolution + bilinear interpolation) to recover original image resolution. ASPP probes an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thus capturing objects as well as image context at multiple scales. It concatenates the responses of the final DCNN layer with fully connected CRFs which improves the localization performance.

2.1.3 DRINet for Medical Image Segmentation (Liang Chen et al., 2018)

Problem: Learn features more distinctively when there are very subtle differences in image attributes such as location, shape, size and intensity.
Key idea: The main motivation of this paper is to better the architecture to distinguish between subtle differences between several categories of the images. The authors point out the success of U-Net architecture and its limitations to scale by adding more layers so that it could learn more representative features. This results in less efficient performance when it comes to differentiating the images based on the intensity, shape, size and location of the image. The authors propose a new architecture comprising three parts - a convolutional layer with dense connections, a convolutional layer with residual inception modules and a pooling layer. The proposed architecture is shown to perform better than the U-Net in three different medical applications.

2.1.4 Skin Cancer Segmentation and Classification with NABLA-N and Inception Recurrent Residual Convolutional Networks (Alom et al., 2019)

Problem: A dermoscopic image segmentation architecture with better feature fusion techniques to provide high image recognition accuracy.
Key idea: The main motivation of this paper is to come up with recent improvements in DCNN based image segmentation architecture which performs dermatologist level disease identification. Major obstacles as per the authors are the varying orientations, illuminations, lighting conditions and several other such inconsistencies in images which causes very low accuracy in image recognition. Major segmentation models such as U-Net, Res-UNet etc. have two basic components - encoder and decoder. The approach they take to solve the above problem of low recognition is by enhancing the ability to decode a unit to produce better and more accurate outputs. For this they propose the NABLA-N Net Model, which produces a better representation of features from the decoding unit by utilizing multiple feature spaces in the deeper layers of an encoding unit by using multiple decoders. These multiple decoders decode the encoding units from different latent spaces. They use Inception Recurrent Residual Convolutional Network for the skin cancer classification which as per them shows better results then other equivalent models.

2.1.5 DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation (Jha et al., 2020)

Problem: Improve performance of U-Net architecture for various segmentation tasks.
Key idea: The main motivation of the authors is to improve performance of the well known U-Net architecture by adding another layer of U-Net architecture stacked on top of each other. They also modify one layer of U-Net by using a pre-trained encoder VGG-19. Adding another layer allows capturing semantic information more efficiently as it makes the network capture more representative features and thus producing better outputs. Further to capture the contextual information within the network they use Atrous Spatial Pyramid Pooling. Based on the experimental results, the architecture shows improvement over U-Net architecture on several datasets.

2.1.6 Towards Automated Melanoma Detection with Deep Learning: Data Purification and Augmentation (Bisla et al., 2019)

Problem: Improve performance of Image segmentation by Data Augmentation.
Key idea: Authors leverage on existing methods, Schmid-Saugeona (2003), to remove occlusions such as hair, rulers etc. Schmid-Saugeona (2003) uses a threshold for the luminance channel of the LUV color space to remove hair. This may have the side-effect of removing dark regions belonging to the lesion. The authors overcome this limitation by using the lesion segmented image and overlaying that with the processed image. Any holes arising from this were closed by using "closing" operation. Next U-Net Ronneberger et al. (2015) architecture was used for lesion segmentation. The authors modified U-Net architecture and experimented with different learning rates to obtain optimal hyperparameters. Typical ML classifiers tend to be biased towards majority classes. Majority of images correspond to benign lesions. This leads the classifier to be biased towards benign classification. Authors use two separate Deep Convolutional Generative Adversarial Networks (DCGANs) to generate synthetic images of melanoma and seborrheic keratosis. The authors ensure the generated images differ from the original dataset by using mean squared error (MSE) to choose the images with least MSE. In the next step, the data is augmented by performing horizontal and vertical flipping of the images.

2.1.7 Melanoma diagnosis using deep learning techniques on dermatoscopic images (Acosta et al., 2021)

Problem: Mask and Region based Convolutional Neural Network (M R_CNN) for creating "Region of Interest" (Bounding Box).
Key Idea: Authors propose a two stage classification method for melanoma detection. First stage uses Mask and Region-based Convolutional Neural Network (M R_CNN) to create a bounding box around the skin lesion. Second stage uses ResNet152 to classify the cropped area. Masks determined by a clinical expert were used to train Mask R_CNN in order to identify the pixels belonging to lesions. The authors tried five different techniques to balance the dataset. Data augmentation techniques experimented are: rotation augmentation (image rotation by 180 degrees) and vertical flip augmentation (pixel reordering by row reversal). Authors present results of their experiments and compare the performance of their model to models reported for the 2017 ISBI Challenge. Authors make a convincing case for using an automated scheme for extracting the region of interest (lesion) in order to improve classifier performance.

2.1.8 Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images (Codella et al., 2016)

Problem: Using non-linear distortions for Data Augmentation.
Key Idea: Authors propose a combination of hand-coded feature extractors, sparse coding techniques, SVMs, Deep residual networks and fully convolutional neural networks to develop ensembles for melanoma segmentation and classification. Authors used a network similar to U-Net Ronneberger et al. (2015) for lesion segmentation. The authors describe the network in

detail and provide motivation for the network topology. The network is trained using six color channels, including Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV) color spaces. Using empirical experimentation authors found improvements in performance on training data using six color channels as opposed to three. Data augmentation is used in each training batch. Images are rotated, flipped, rescaled, shifted, cropped, and nonlinear distortions. Nonlinear distortions appropriately model the variation of soft-tissue biological structures. Authors show performance on segmentation tasks, which places them in the second rank for 2016 ISBI challenge

2.2 Transformer-based Approach

2.2.1 An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Dosovitskiy et al., 2020)

Problem: It evaluates the use of pure transformers based approach (without using CNN) for image classification.
Key Idea: Transformers have been widely used in NLP tasks to learn patterns from sequences, however they suffer from quadratic complexity. Due to this limitation, they are not suitable for longer input sequences. To solve this problem for images, the authors proposed an approach of dividing images into patches of 16 by 16 and then treating each patch as a single input to the transformer. The flattened patch vector is then fed as an input to the transformer architecture.

2.2.2 TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (Chen et al., 2021)

Problem: Achieving state-of-the-performance in biomedical images using Transformer-based architecture.
Key idea: The paper combines the merits from U-Net and Transformers to develop an architecture which gives local as well as global attention. It uses a CNN-Transformer Hybrid as an encoder. The CNN layer output is a 1 by 1 patch which is fed into the visual transformer to generate a latent feature representation. This representation is passed through a cascaded upsampler which is similar to the upsampling step of the U-Net. It also uses skip-connections at different layers to leverage the localized attention from the CNN feature map in the decoding path.

2.2.3 Convolution-Free Medical Image Segmentation using Transformers (Karimi et al., 2021)

Problem: Medical image segmentation with a convolution-free model based entirely on self-attention.
Key idea: CNNs have been shown to be highly effective in various computer vision problems. Convolutions bring important properties such as sparse interactions or weight sharing which form a strong and useful inductive bias (ability to generalize) for vision tasks. However, medical image analysis applications come with additional challenges like a 3D structure (MRI) or a small

number of labeled images. To tackle those challenges the authors propose a new model based entirely on self-attention. They build on the insight of the vision transformer and split the image into (3D) patches which are then flattened, embedded and concatenated with a positional encoding (optional). This representation is fed into an encoder with a multi-headed self-attention which produces the final segmentation mask for the center patch of the image.
They also show that pre-training the proposed model on large corpora of unlabeled images gains a significant performance boost when labeled training data is small (5-15 examples).

2.2.4 Medical Transformer: Gated Axial-Attention for Medical Image Segmentation (Valanarasu et al., 2021)

Problem: Application of Transformer-based solutions for medical image segmentation with small datasets.
Key idea: The lack of understanding of long-range dependencies of Deep Convolutional Neural Networks motivated the authors of this paper to apply Transformer-based architectures with self-attention mechanism for medical image segmentation tasks. The availability of only a small annotated dataset makes it difficult to efficiently train transformers. To address this issue the authors propose the Medical Transformer (encoder-decoder architecture) which adds an additional control mechanism called Gated Axial-Attention in the self-attention module (encoder) as well as a Local-Global training strategy. This strategy helps the model to operate on the whole image resolution where it learns high-level features by modelling long-range dependencies (global), and on image patches which helps to focus on finer local features at the same time (local).

3 Compare and Contrast

This chapter compares the presented papers. First, we look at what they have in common and where they contrast in more detail. Then we summarize what aspects of machine learning theory the authors use for their work.

3.1 Article Comparison

3.1.1 CNN-based Approach

(Ronneberger et al., 2015) was a seminal paper which proposed an architecture which was against the larger consent that a successful training required thousands of images. U-Net architecture comprises two parts - the first, where features are learned and the other where the learning is applied for segmentation. It preserves the feature resolution during upsampling by adding skip connections. This work is often used by other authors as a baseline in their work to come up with new, better performing architectures. The other works which we analyze in this review also are focussed on improving U-Net, using some improvement in the original architecture so that it can detect subtle differences in features in images in different categories. For example, the Double-UNet authors (Jha et al., 2020) use two layers of U-Net along with

Atrous Spatial Pyramid pooling whereas DeepLab (Chen et al., 2016) tackles the problem of reduced feature resolution by using atrous convolution. Compared to the work based on U-Net, (Alom et al., 2019) present a new segmentation model, the Nabla-N architecture,which proposes to use better feature fusion techniques in the decoding layer to achieve a better feature representation. Their experimental results show higher accuracy on segmentation tasks, reaching around 87% testing accuracy for dermoscopic classification of skin cancer on ISIC 2018 dataset. (Bisla et al., 2019) propose a Data Augmentation technique to improve the performance of U-Net architecture. (Acosta et al., 2021) utilize a different approach for image segmentation, they use a Mask and Region based Convolutional Neural Network (M R_CNN) for creating “Region of Interest”. (Codella et al., 2016) propose a data augmentation technique to improve performance on segmentation performance. They successfully show the merits of using non-linear distortions to improve segmentation performance.

3.1.2 Transformer-based Approach

Vision Transformers (Dosovitskiy et al., 2020) is a seminal paper as it classifies images completely using transformer architecture and removes the need of CNN in vision. CNN architecture only captures local attention while the Transformer-based vision architecture captures global attention over every pixel of the image. It opens the door for a lot of interesting research and ideas. The paper also shows that it takes less time to train the Vision Transformer model (ViT) as compared to ResNet-based architecture.

TransUNet (Chen et al., 2021) is inspired by ViT (Dosovitskiy et al., 2020) and combines the U-Net and vision transformer based architecture by using cascading CNN’s and transformers for downsampling. It then uses skip connections in the upsampling path, hence persevering localized as well as global information.

Medical Transformer (Valanarasu et al., 2021) extends the work of (Dosovitskiy et al., 2020) by adding the Local-Global strategy to capture the local, finer image features and to model the long-range dependencies present in an image. They also add an additional control mechanism in the self-attention module called Gaten Axial-Attention with four control gates defined as learnable parameters. Compared to (Chen et al., 2021) which depends on pretrained weights obtained by training on a large image corpus, the authors of (Valanarasu et al., 2021) explore the feasibility of applying transformers working on only self-attention mechanisms as an encoder. In medical applications pretraining on large large-scale image datasets becomes problematic as there is often only a small amount of labeled images available as data annotation is very expensive and requires expert knowledge.

Similar to (Valanarasu et al., 2021) the work of (Karimi et al., 2021) focuses on neural networks for medical image segmentation but extends to the 3D space. The authors show that their proposed model can be effectively trained with only 20-200 labeled images and also propose methods that can improve the segmentation accuracy when large corpora of unlabeled training images are available. In contrast to (Valanarasu et al., 2021) and (Dosovitskiy et al., 2020) that use some convolutional layer for feature extraction, (Karimi et al., 2021) relinquish all convolutional layers and explore self attention-based deep neural networks only.

3.2 Machine Learning Theory

The following table 1 summarizes the machine learning theory aspects applied in the presented papers:

Paper	Approach	Model	Metrics	Datasets
Ronneberger et al., 2015	CNN	U-Net	IoU (Intersection over Union)	PhC-U373 (Glioblastoma) (35 partially annotated training images) DIC-HeLa (HeLa Cells) (20 partially annotated training images)
Chen et al., 2016	CNN	DeepLab	IoU	PASCAL VOC 2012 (train: 10582 images, test: 1456 images) Cityscapes (train: 2975 images, val: 500 images, test: 1525 images)
Chen et al., 2018	CNN	DRINet	DSC, SE, Hausdorff95 distance, Precision, Recall	Local PACS databases (train: 500 images, validation: 281 images, test data: obtained separately from 32 subjects)
Alom et al., 2019	CNN	NABLA-N	IoU, Dice Coefficient, F1, Precision, Recall	ISIC 2018 (train & validate: 2100 images, test: 494 images)

Jha et al., 2020	CNN	DoubleU-Net	Sorensen Dice Coefficient (DSC), mean IoU	2015 MICCAI sub-challenge on automatic polyp detection dataset (808 images) CVC-Clinic DB (612 images) Lesion boundary segmentation challenge (2594 images) 2018 Data science bowl (670 images)
Bisla et al., 2019	CNN	U-Net with Data Augmentation	Jaccard Index	ISIC-2017, ISIC-2018 (1875 Images)
Acosta et al., 2021	CNN	Mask and Region-based Convolutional Neural Network (Mask R_CNN)	Receiver Operating Characteristic (ROC) space	ISIC 2017 (Train: 1995, Validation: 149, Test: 598)
Codella et al., 2016	CNN	U-Net with Data Augmentation with non-linear features	Jaccard index and pixel-wise accuracy	ISIC 2016 Train: 900, Test: 379
Chen et al., 2021	Transformer	TransUNet	DSC (Dice similarity coefficient)	Synapse multi-organ CT dataset (train: 1930 images , test: 550 images)

Karimi et al., 2021	Transformer	(no specific name)	DSC (Dice similarity coefficient)	Brain cortical plate (train: 18 images, test: 9 images) Pancreas (train: 231 images, test: 50 images) Hippocampus (train: 220 images, test: 40 images)
Valanarasu et al., 2021	Transformer	MedT	F1 (ablation study), IoU	Brain anatomy segmentation (ultrasound) (1300 2D US scans for training, 329 for testing) Gland segmentation (train: 85 images, test: 80 images) MoNuSeg (train: 30 images, test: 14images)

Table 1: Comparison of machine learning theory aspects.

4 Future Work

This last chapter discusses the future work as well as some of the open questions from the authors' research work. We use those as an inspiration and starting point for our final project.

4.1 Open Questions

The work of (Ronneberger et al., 2015) could be possibly applied to other medical segmentation tasks like CT or MRI scans as well as non-medical applications like cityscape or clothes segmentation in order to obtain a broader and better understanding of the performance of the model in other domains where more annotated images might be available for training.

The authors of the (Karimi et al., 2021) paper expect that the proposed architecture should be effective on other medical image analysis tasks such as anomaly detection and classification and leave this thesis for future work.

The authors of (Chen et al, 2018) paper on DRINet point to a limitation that the increase of the growth rate results in many more parameters which results in many more parameters and causes difficulty in training and slower testing of the model. Future work can attempt to simplify the network structure to improve the performance of the architecture.

The authors of (Codella et al., 2016) paper propose nonlinear image warping technique for Data Augmentation. This scheme might be useful for other classification models. Additional techniques such as using residual convolutional layers for semantic segmentation, meta-learning or boosting for selection of network ensembles to perform segmentation may improve performance.

4.2 The Segue

The covered literature forms a broad foundation for our final project. Based on the presented work we plan to use some of the CNN-based (e.g. U-Net) and Transformer-based (e.g. TransUNet, MedT) models for our proposed task of melanoma image segmentation and extend the presented work with new experiments.

5 References

Md. Zahangir Alom, Theus H. Aspiras, Tarek M. Taha, and Vijayan K. Asari. 2019. Skin cancer segmentation and classification with NABLA-N and inception recurrent residual convolutional networks. *CoRR*, abs/1904.11126.

Devansh Bisla, Anna Choromanska, Jennifer A. Stein, David Polsky, and Russell Berman. 2019. Towards automated melanoma detection with deep learning: Data purification and augmentation.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueck- ert. 2018. Drinet for medical image segmentation. *IEEE Transactions on Medical Imaging*, 37(11):2453–2462.

Liang-Chieh Chen, George Papandreou, Iasonas Kokki- nos, Kevin Murphy, and Alan L. Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915.

Noel C. F. Codella, Quoc-Bao Nguyen, Sharath Pankanti, David Gutman, Brian Helba, Allan Halpern, and John R. Smith. 2016. Deep learning ensembles for melanoma recognition in dermoscopy images. *CoRR*, abs/1610.04662.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Debesh Jha, Michael A. Riegler, Dag Johansen, Pål Halvorsen and Håvard D. Johansen 2020. Double u-net: A deep convolutional neural network for medical image segmentation.

Davood Karimi, Serge Vasylechko, and Ali Gholipour. 2021. Convolution-free medical image segmentation using transformers.

Maria Begonya Garcia-Zapirain Mario Fernando Jojoa Acosta, Liesle Yail Caballero Tovar1 and Win- ston Spencer Percybrooks. 2021. Melanoma diagnosis using deep learning techniques on dermatoscopic images. *BMC Medical Imaging*, arXiv:1503.06733.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Guillodb J. Thirana J. P. Schmid-Saugeona, P. 2003. Towards a computer-aided diagnosis system for pig- mented skin lesions. *Journal of the Computerized Medical Imaging Society*, pages 67–78.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Haci- haliloglu, and Vishal M. Patel. 2021. Medical transformer: Gated axial-attention for medical image segmentation.