2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

UNIVERSAL MULTI-MODAL DEEP NETWORK FOR CLASSIFICATION AND SEGMENTATION OF MEDICAL IMAGES

Ahmed Harouni, Alexandros Karargyris, Mohammadreza Negahdar, David Beymer, Tanveer Syeda-Mahmood

IBM Research-Almaden, San Jose, CA, USA

ABSTRACT

Medical image processing algorithms have traditionally focused on a specific problem or disease per modality. This approach has continued with the wide-spread adoption of deep learning in the last 5 years. Building a system with multiple neural networks and different specialized image processing algorithms is a challenge as each network requires a lot of memory and is computationally heavy. More importantly, cascading multiple networks propagates errors from one stage to another reducing overall system accuracy. In this work, we propose a single universal network that can: 1) segment different organs across different modalities, and 2) solve both segmentation and classification problems simultaneously. We compare our approach with traditional segmentation network for each modality. Our results showed modality/viewpoint classification accuracy of 99% and average dice score of 0.89 for segmentation accuracy. The proposed network can be further developed to include segmentation of more organs and disease classification.

Index Terms— Deep network, Segmentation, Unet, classification, multi-modality, universal network

1. INTRODUCTION

Over the last decade Deep Learning has become a prominent area of research in machine learning due to recent advances in theory (solvers and optimizers) and infrastructure (larger-memory and faster graphic processing units). Convolutional Neural Networks (CNNs) [1, 2] have gained tremendous popularity within the computer vision community because of their ability to automatically capture high level representations of raw images. This approach has elevated the need to hand crafted features customized for each problem. CNNs have shown state-of-the-art results in image classification, object detection and segmentation. It is because of these reasons that CNNs have taken over the medical image analysis field in the past few years helping achieve great improvements in disease classification, image registration and anatomy segmentation [3].

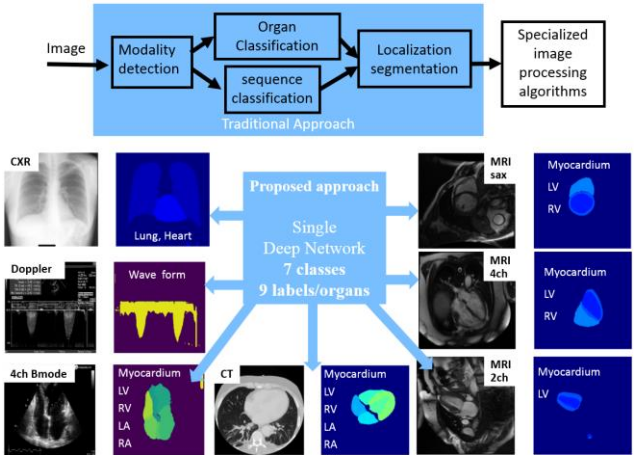

Figure 1: Top: Traditional architecture tackling one problem at a time. Bottom: our proposed network to both classify different modalities with different viewpoints (X-ray, CT, Ultra sound, 2 chamber MRI, 4 chamber MRI, short axis MRI) as well as segment different structures as Lung, heart, Doppler wave form, Myocarduim (Myo), left ventrical (LV), right ventrical (RV) ,left atrium (LA), and right atrium (RA).

However, in order to properly train Deep Learning systems, such as CNNs, a large number of examples are required to tune a large number of parameters. In medical image analysis this problem is very critical due to a) the cost of collecting medical images, b) the regulatory constraints of acquiring medical images, and c) the cost and time of annotation (i.e. ground truthing) by clinicians. Litjens *etal* [3] surveyed more than 280 papers, where the main approach applied by researchers is to train a Deep Learning system per medical modality/view to achieve a specific task (e.g. heart ventricle segmentation in MRI). This approach, however, subsequently raises an important technical issue in a radiology setting: it requires a large number of deep learning networks be loaded in the memory each one addressing a specific task. This makes scaling almost impossible given the large number of anatomies and modalities found in radiology. Finally, building one network per modality/view per task requires a lot of examples per modality/view because of the large size of network

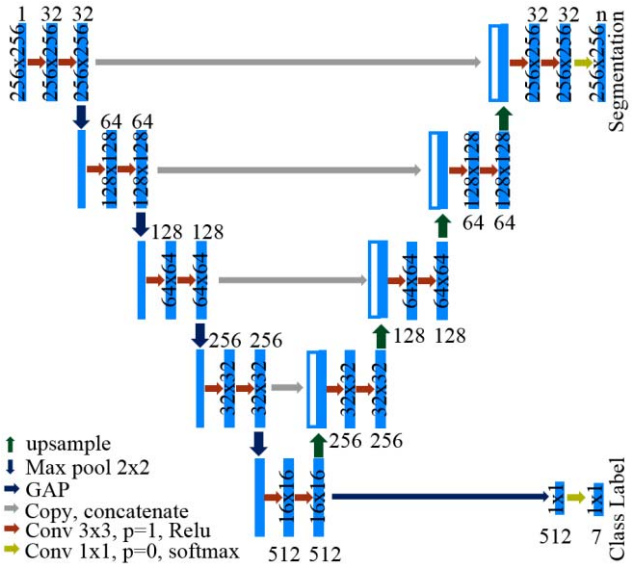

Figure 2: Single network architecture with two heads for segmentation and classification. Last layer of the segmentation has *n* filters corresponding to *n* structures depending on the dataset.

parameters. However, if the network was decoupled from the modality/view constraint then examples from various modalities/views could be used together to train this single network. This approach would allow for a more efficient solution since the network could be trained using even a few examples acquired from a new modality/view. Figure 1 shows a traditional approach of networks connected in sequence, beginning with modality classifier, technique/sequence classifier, view point detection, organ localization, and ending with an image processing pipeline to extract measurements, detect diseases, and generate reports. For example, modality classification would classify images into different modalities as MRI, CT, US, X-ray, ECG, EEG; An MRI sequence classification would classify images according to sequence type as SSFP, T1W, T2W, FLIAR, inversion recovery, etc. One major disadvantage of such system is that errors propagate from one level to the next, deteriorating the overall accuracy of the system.

In this work, we propose: a) combining data from different modalities and viewpoints to train a single network, b) training a single universal network for segmentation and classification tasks as shown in Fig. [1]. This method could be applied to any segmentation network as in [5-7].

2. METHODS

Our network architecture is based on Unet architecture [2] with two output heads, one for segmentation and the other for classification as shown in Fig [2]. The network consists of a concatenating path and an expanding path, the classification and segmentation outputs are at the end of the concatenating and expanding paths respectively. The concatenating path is 4 levels deep with 2x2 pooling between each level, while the expanding path is connected by an upsampling filter with 2x2 kernel. All convolutions have a kernel size of 3x3, stride=1, pad =1 followed by Relu. Padding maintains the size fixed before and after convolution. Each level is composed of 2 convolutions back to back. The last layer of the segmentation path is *n* filters (*n* is the number of segmented structures) of 1x1 convolution followed by a softmax, which gives the probabilities of the segmentation labels. Skip connections between the layers are used to avoid vanishing gradient problem. For classification task, we added a global average pooling (GAP) layer at the end of the concatenating path, followed by a convolution layer with 7 filters corresponding to the 7 classes (see table [1]), and finally a softmax layer to produce the class probabilities. Since the network is optimizing two loss layer (segmentation and classification), we define α as the ratio between the weight of the classification loss and the segmentation loss.

Modality	Class label	Segmentation label	# Images	
			Trn	Val
X-ray	CXR	Lung, Heart	200	46
MRI	SAX	Myo, LV, RV	544	260
	2Ch	Myo, LV	27	21
	4Ch	Myo, LV, RV	35	19
CT	CT	Myo, LV, RV, LA, RA	1535	410
Ultra Sound	Bmod	Myo, LV, RV, LA, RA	40	10
	Dop	Doppler wave form	400	250
Total			**2781**	**1016**

Table 1: Dataset distribution for training set and validation set for each classification and segmentation task. Classification task spans 7 classes covering 4 modalities and 3 viewpoints of MRI. Segmentation task spans 8 structures besides the background.

3. DATASET

Our dataset is composed of multi modal cardiac images that includes: MRI, CT, chest X-ray (CXR) and Ultrasound as shown in table [1]. X-ray images are frontal Chest X-ray (CXR) obtained from publicly available dataset. MRI images are from Sunny brook dataset, which is steady state free precision (SSFP) sequences for 45 patients. All patients had multi-slice short axis (SAX), while some patients have two chamber (2Ch), and four chamber (4Ch) views. CT images were for chest CT exams for 20 patients covering the chest area, we only used slices which contains the heart. Echo Ultrasound data consists of four chamber B-mode view and doppler obtained from a collaborating hospital. This divides the dataset into 7 classes spanning 4 modalities and

Network	Training Dataset	Task		*n*	Segmentation dice score									Classification Accuracy
		Cls	Seg		BG	Lung	Heart	LV	Myo	RV	LA	RA	Doppler	
Unet-CXR	X-ray		x	3	0.97	**0.95**	**0.93**	-	-	-	-	-	-	-
Unet-MRI	MRI		x	4	0.99	-	-	**0.90**	0.79	0.85	-	-	-	-
Unet-CT	CT		x	6	0.99	-	-	0.73	**0.85**	**0.86**	0.87	0.89	-	-
Unet-US	Ultra sound		x	7	0.94	-	-	0.79	0.79	0.81	0.74	0.84	0.75	-
Unet-AllMode	All modalities		x	9	0.99	**0.95**	0.86	0.83	0.84	**0.87**	0.87	0.89	**0.84**	-
Unet-Cls	All modalities	x	x	9	0.99	0.94	**0.89**	0.83	**0.85**	**0.86**	**0.88**	**0.90**	0.83	**99.5%**
AlexNet	All modalities	x		-	-	-	-	-	-	-	-	-	-	98%

Table 2: Dice score of different structures and classification accuracy for different network architectures trained on different datasets for segmentation (Seg) and classification (Cls) tasks. *n* refers to the number of segmented structures and number of the output segmentation layers. Bold font highlights highest score per structure.

3 different MRI orientations. A radiologist with 10+ years of experience segmented all datasets. Segmented structures were: lungs and heart for CXR images; myocardium muscle (Myo), left ventricle (LV), right ventricle (RV), left atrium (LA) and right atrium (RA) for CT images; Myo, LV, RV for MR images; Myo, LV, RV, LA, RA and Doppler wave form for Ultra sound images, and Background (BG). Data was split at a patient level into training (~65%) and validation dataset (~35%) as shown in Table [1].

4. EXPERIMENTS

All network architectures were implemented in Caffe. All images were resized to 256x256 resolution. Data augmentation were performed on the fly while training to include selecting a random patch of 128x128 then applying horizontal flips, rotations (-10˚ to +10˚) and scaling (0.8 to 1.2). Networks were trained for 150 epochs with batch size of 16. Stochastic optimization algorithm Adam was used with base learning rate =0.0001, momentum =0.9. Sigmoid decay function (step =50%, gama =0.1) was used to update the learning rate between epochs.

In order to show how our proposed universal architecture compared against the state-of-the-art networks, we trained multiple networks, one for each modality. We will refer to Unet-CXR, Unet-MRI, Unet-CT, Unet-US as a typical Unet architecture trained and validated only on CXR, MRI, CT, and Ultra-sound images respectively as shown in table [2]. All modalities were combined to train and validate a single Unet referred to as Unet-AllMode. All Unet networks had same parameters as shown in Fig. [2] except for the number of output layers (*n*) which is equal to the number of segmentation labels as shown in table [2]. We will refer to our network architecture as Unet-Cls to indicate that it provides segmentation as well as classification. We trained Unet-Cls using all modalities with different values of α. In order to mimic the traditional approach (see Fig. [1]), we trained AlexNet [1] for the classification of the 7 classes shown in table [1]. Dice coefficient was used as a measure of segmentation while confusion matrix and accuracy rates were calculated for classification tasks.

5. RESULTS

Unet-Cls were trained using α values of 0.1, 0.3, 0.5, 1.0 1.5, 2.0. Different values of α had small effect of average dice score (less than 0.03). Classification accuracy for our proposed network were above 99% for different values of alpha compared to 98% accuracy for AlexNet architecture. Unet-Cls with α of 0.1 yield best balance between accuracy and dice score. Table [2] shows dice score for each anatomical structure segmented by different network architectures. The lowest dice score was 0.73, while the largest dice score among difference architectures was 0.07. Figure 3 shows misclassified regions inferred by single modality networks (see white arrows). This clearly indicates the need for a modality classification network before the segmentation network as performed in conventional approaches (see top Fig. [1]). This cascaded architecture adds more complexity to the system and decreases the overall accuracy. Figure 4 shows segmentation results of different modalities for different networks architectures. Unet-Cls is able to avoid segmentation errors as indicated by white arrows in Fig. [4].

6. DISCUSSION

The best segmentation results were expected to be obtained from specialized network (trained on single modality) since each of those networks have only seen these type of images. However, by examining table [2] we can see that networks trained on single modalities are within 0.02 dice values from the Unet-AllMode, with the exception of the LV dice value that was 0.07 points below for the Unet-AllMode. This may be due to the fact that LV segmentation for CT and US is much harder than MRI images as indicated by dice scores of Unet-CT (0.73) and Unet-US (0.79) vs Unet-MRI (0.9). Interestingly, Unet-MRI and Unet-CT, which were trained on MRI and CT images respectively, manage to correctly locate and partially segment the LV, LA and RV in MR and CT images (see red arrows in Fig. [3]). These results were our motivation to design a single universal network trained on multiple modalities. AlexNet and Unet-Cls (with different

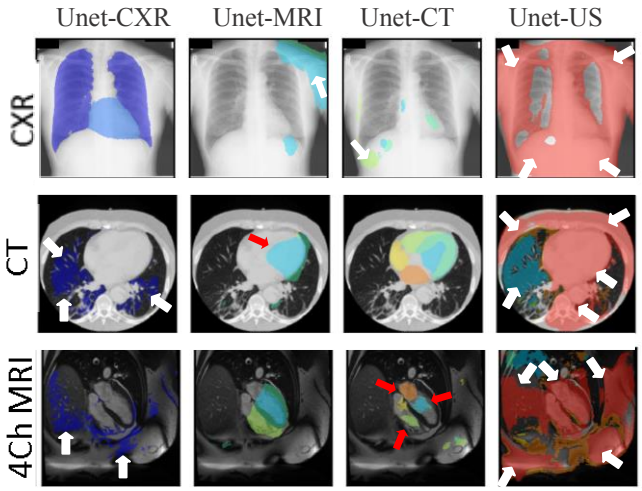

Figure 3: Segmentation results for single modality trained networks Unet-CXR, Unet-MRI, Unet-CT, and Unet-US for chest X-ray (CXR), CT and 4 chamber MR images. White arrows points to mis-segmentation by Unet-CXR, Unet-US and Unet-MRI networks indicating the need for a classification network. Red arrows points to Unet-MRI and Unet-CT network segmenting parts of the ventricles in CT and MRI images respectively although they were not in their perspective training data.

values of α) all revealed very high accuracy over 98%. This may indicate that the classification problem is relatively simple, which explains why Unet-Cls didn't get a dice score boost from the classification head. Table 2 shows that Unet-Allmode and Unet-Cls have similar dice score for all structures, however, Unet-Cls provides the classification for free. More importantly, Unet-Cls has half the inference time compared to other architectures since the segmentation and classification are generated from the same network instead of cascaded networks. One minor disadvantages is that Unet-Allmode and Unet-Cls required longer training time as the training dataset was larger.

7. CONCLUSION AND FUTURE WORK

In this work, we proposed using a single universal network architecture to classify and segment multi-modal medical images. To the best of our knowledge this study is the first effort in the medical domain to combine such diverse modalities with different structures. Our results show that combining different modalities yields similar and sometimes better results for shared structures than using separate architectures for each modality. Since our architecture is a single network, it occupies less memory and resources by using fraction of parameters (compared to multiple single modality networks). In addition, it avoids error propagation compared to traditional approach shown in Fig. [1].

Future work includes extending our dataset with other MRI sequences (T1W, T2W, delayed enhancement) as well as different MRI orientations (i.e. 3 chamber and axial

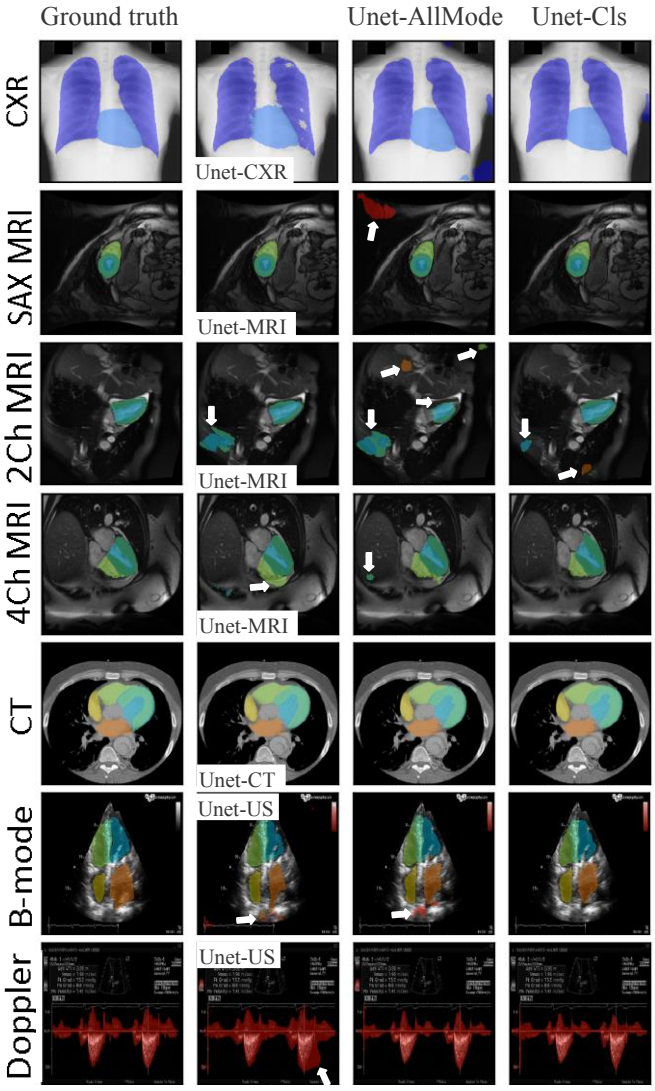

Figure 4: Segmentation results for different modalities: chest x-ray (CXR), 2 chamber MRI, 4 chamber MRI, short axis (SAX) MRI, CT, 4 Chamber B-mode echo, and Doppler. Columns left to right show ground truth segmentation, Unet trained on single modality data, Unet-AllMode, Unet-Cls segmentation results. White arrow points to mis-segmentation by single modality Unet and Unet-AllMode that is avoided by Unet-Cls.

views). We also plan to extend this to other echo views (i.e. 2, 3, 5 chambers) as well as different organs (brain and abdominal structures). Finally, we want to extend the same architecture in disease detection.

8. REFERENCES

[1] Krizhevsky, A., Sutskever, I., Hinton, G, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS, pp. 1097-1105, 2012.

[2] Ronneberger, O., Fischer, P., Brox, T., “U-Net: convolutional networks for biomedical image segmentation” MICCAI LNCS, vol. 9351, pp. 234–241, 2015.

[3] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, *etal*. , “A survey on deep learning in medical image analysis”, In Medical Image Analysis, vol. 42, Pages 60-88, 2017.

[4] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., “Gradient-based learning applied to document recognition”, Proceedings of the IEEE vol.86, pp. 2278–2324, 1998.

[5] Long, J., Shelhamer, E., & Darrell, T. “Fully convolutional networks for semantic segmentation” IEEE Conference on Computer Vision and Pattern Recognition pp. 3431-3440, 2015.

[6] Cicek, O., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., “3D U-Net: Learning dense volumetric segmentation from sparse annotation”, MICCAI, Vol. 9901, pp. 424–432, 2016.

[7] Milletari, F., Navab, N., Ahmadi, S.-A., “V-Net: Fully convolu-tional neural networks for volumetric medical image segmentation”, 2016.