UNET-Based Multi-Task Architecture for Brain Lesion Segmentation

1st Ava Assadi Abolvardi
dept. Computing
Macquarie University
Sydney, Australia
ava.assadi-abolvardi@hdr.mq.edu.au

2nd Len Hamey
dept. Computing
Macquarie University
Sydney, Australia
len.hamey@mq.edu.au

3rd Kevin Ho-Shon
dept. Clinical Medicine
Macquarie University
Sydney, Australia
kevin.ho-shon@mq.edu.au

***Abstract*— Image segmentation is the task of extracting the region of interest in images and is one of the main applications of computer vision in the medical domain. Like other computer vision tasks, deep learning is the main solution to image segmentation problems. Deep learning methods are data-hungry and need a huge amount of data for training. On the other side, data shortage is always a problem, especially in the medical domain. Multi-task learning is a technique which helps the deep model to learn better representation from data distribution by introducing related auxiliary tasks. In this study, we investigate a research question to whether it is better to provide this auxiliary information as an input to the network, or is it better to use this task and design a multi-output network. Our findings suggest that however, the multi-output manner improves the overall performance, but the best result achieves when this extra information serves as auxiliary input information.**

Keywords*— Deep Learning, Multi-task learning, image segmentation, computer vision.

I. Introduction

Image segmentation is one of the computer vision tasks which extracts the region of interest [1, 2]. Image segmentation has several applications in the medical domain, from organ segmentation to tumour segmentation in medical images [3, 4, 5]. Deep neural networks, outperform other techniques in this field and have become the number one solution to such problems [6]. The disadvantage of deep models is that they need lots of annotated data to learn from. Shortage of available annotated data is a big challenge in the medical area since data annotation is a tedious task which needs expertise [7]. The concept of multi-task learning is to introduce a related auxiliary task to the network and design the network to learn both tasks simultaneously [8]. This auxiliary task, not only brings extra data with itself, which is crucial for data shortage challenge, but also helps the network to generalize better and learns a more powerful representation from the data. Intuitively, the idea of multi-task learning is inspired by how humans use the knowledge acquired from learning other related tasks in the target task to increase the performance. It is proven that when making a classifier to do another auxiliary task, the overall performance on the first task will improve [9]. In this study, our focus is on Multi Sclerosis (MS) lesion segmentation. Multiple Sclerosis is an autoimmune disorder which eats away the protective covering of the nerves, causing the appearance of lesions in the brain [10]. MS is also known as white matter disease since they only appear in the white matter. One way of using this information is to design a multi-task architecture in which not only the network is trained to segment the lesions, but also is trained to segment brain tissues as well. Brain tissue segmentation is the task of segmenting the brain MRI into brain tissues which are white matter, grey matter and CSF [11]. Following image shows both lesion segmentation and brain tissue segmentation tasks.

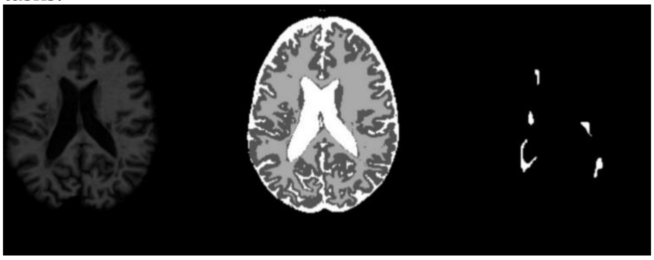

Fig. 1. Brain MR Image (Left), Brain Tissue Segmentation (Center), Multiple Sclerosis Lesion Segmentation (Right).

In figure 1, the left image is the brain MRI, the middle image is the mask of brain tissues, and the image in the right is the mask of MS lesions in the brain.

As mentioned before, since MS lesions only occur in the white matter, one way of using this information is to design multi-output network architecture, in which the network tries to segment both lesions and brain tissues at the same time. The intuition is that, when the network tries to find a mapping between input and two outputs, it also learns the correlation between white matter and the lesions. Another way of using such information is to feed this extra information directly to the network as input. In the latter form, the architecture of our deep model becomes a multi-input network. The aim of this study is to answer this research question, to whether to use extra information as an input or output?

In the following, we first will discuss the background of multi-task learning.

II. Background

Multi-task learning is known under different names such as joint-learning, learning to learn or learning with an auxiliary task, but in fact, when you train your network regarding more than one loss function, the network is considered as multi-task learning [12]. One of the main categorizations of multi-task learning divides the techniques into hard-parameter sharing [13] and soft parameter sharing [14] categories.

A. Hard Parameter Sharing

The most popular deep multi-task learning method is hard parameter sharing in which, hidden layers are shared between all tasks while having a task-specific output for each individual task.

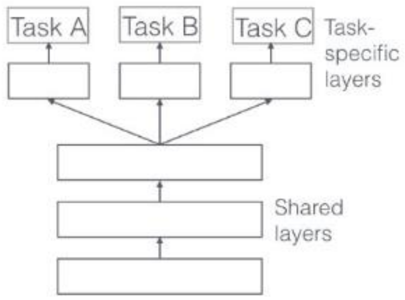

Fig. 2. Hard Parameter Sharing Architecture

Two main architectures that were introduced in this category are Deep Relationship Network and Fully-Adaptive Feature Sharing. Deep Relationship Network was proposed by Long and Wang [15] and the main idea of is similar to hard parameter sharing, with having shared convolutional layers and task-specific fully connected layers. They also used prior matrices in the model to learn relationships between tasks. Fully-Adaptive Feature Sharing architecture starts with a Hard Sharing network and greedily groups similar tasks based on criteria. This network was introduced by Lu [16] in 2016.

B. Soft Parameter Sharing

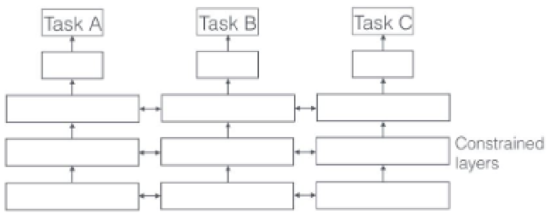

Fig. 3. Soft Parameter Sharing Architecture

Unlike Hard Parameter Sharing, in Soft Sharing, each task has its own model, but the parameters are trained in a way to be as similar as possible. In the case of having exactly the same parameters, the network is identical to Hard Sharing network.

Cross-Stitch networks, introduced by Misra [17] in 2016 is one of the main architectures in multi-task learning era. Cross-Stitch Networks are based upon Soft Parameter Sharing and have separate models for each task. The network is trained to learn how the knowledge of one task can leverage the other task by learning a linear combination of the output of previous layers which is called stitch module. Sluice was introduced in 2017 [18] as an improvement to Cross-Stitch Networks. This model learns what layers should be shared as well as appropriate amount of sharing. Another novelty of this work is that the model is able to learn the appropriate relative weights of the different task losses.

III. Related Works

In recent years, multi-task learning has gained attention in medical domain [19]. Lots of multi-task convolutional neural networks have been proposed. One of the earliest works in this field was done by Moeskops et al in 2016 [20]. In their proposed architecture, they trained a single CNN for three different segmentation task. The network is trained to classify tasks prior to segmentation. In 2017 Wufeng Xue [21], proposed a fully convolutional deep multi task network for full cardiac left ventricle quantification. Proposed method consists of one CNN and two RNNs. Since our focus is on UNET-based multi task learning architectures, we will discuss these methods in this area.

In 2018, Liang Cao [22] proposed a UNET-based multi task network which is trained for segmentation and regression tasks. Segmentation part is trained to extract hippocampus, and regression part calculates Mini-mental state examination (MMSE) scores for subjects.

Xulei Yang [23] also proposed a UNET-based multi task network which segments lesions in skin images as well as classification of skin lesion types. Since skin images are 2D, their proposed method is a 2D version of UNET.

Toan Duc Bui [24] proposed a multi task UNET for Neonatal Brain Segmentation. This network has two paths, one path is for segmentation and the other one is designed to calculate regression. Segmentation path is a 3D-UNET, and the regression path is a convolutional network which shares the encoder part with segmentation path. Following image shows this architecture.

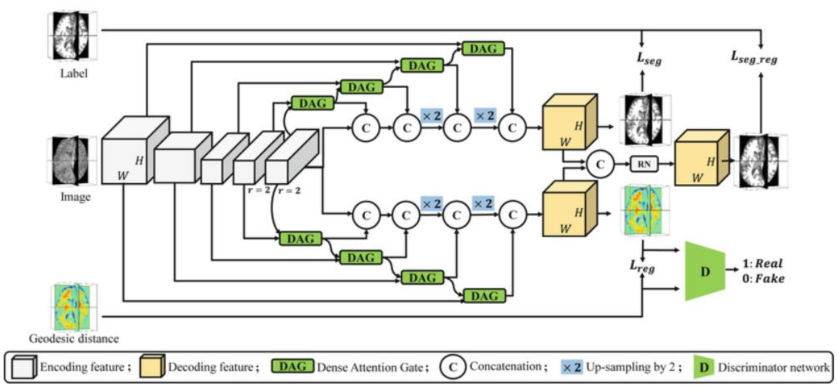

Fig. 4. Multi-Task U-Net Architecture Propsed by Toan Duc Bui.

Most of the multi-task architectures were designed so that it does segmentation in one path and classification or regression in another path. Our aim in this study is to propose a model for two different segmentation tasks. The main difference between our work and other methods, is that our loss function should be able to give good feedback to the network, so the model can segment both tasks efficiently. In the following section we discuss our proposed model.

IV. Proposed Method

In order to compare two strategies, we designed two different architectures; one for multi-input network and one for multi-output network. Since our target task is to segment brain lesions, both of our networks are based on 3D-UNET. UNET is a fully convolutional network which was designed for medical segmentation tasks [25]. In fully convolutional networks, last layer is also a convolutional layer which aims at reconstructing the image, by performing up-convolution. UNET has two paths, the first one is called contracting path which encodes the image, and the expanding path which decodes the image based on the features that it learned during encoding part. The model also uses skip connections [26], also known as residual connections, between the down-sampling and the up-sampling paths. These residual connections will provide local information and are proved to be helpful for the algorithm convergence. UNET, was first designed to deal with 2D images, but later versions [27] were able to deal with 3D image sets as well. Figure 5 shows the UNET architecture.

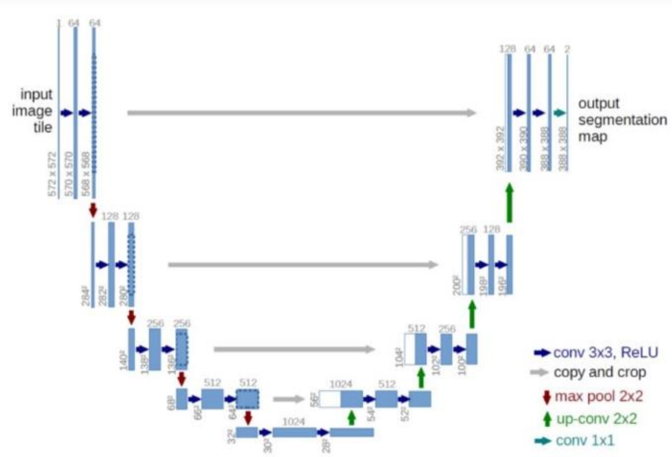

Fig. 5. U-Net Architecture

For our multi-input architecture, we used multi-modal 3D-UNET and passed the each brain tissue mask as a different modality to the network. On the other side for multi-output architecture, we combined hard-parameter sharing with 3D-UNET, and designed a multi-output UNET. Following figure, shows the architecture for our multi-output proposed method.

Figure 6 shows that all the layers are shared except for the last layer in expanding path. There are two different final layers each for one task.

We also customized a loss function for our proposed architecture. Since both of our tasks are segmentation, we introduced a dice-coefficient based loss function. Our loss function has two parts; first term is weighted dice loss [28], and the second term is true positive rate [29]. The reason that we use weighted dice loss is the ability of this function at handling class imbalance [30].

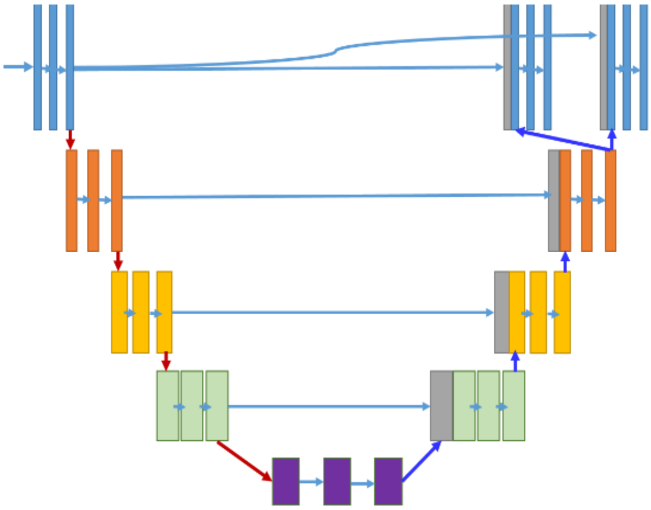

Fig. 6. Our Proposed Multi-Output Architecture

Weighted dice loss can be calculated using following equation:

$$Weighted\ Dice\ (M_R, M_A) = 1 - 2\frac{w|M_R \cap M_A|}{w|M_R| + |M_A|} \quad (1)$$

$$Where\ w = 1/(M_R)^2$$

The weight, in weighted dice loss, corrects the contribution of each label by the inverse of its volume [28]. By doing so, the correlation between region size and dice score will decrease.

As mentioned before, dice loss is beneficial to our both tasks, but is not enough for our main task which is brain lesion segmentation. In the process of extracting lesions of the brain, there's another metric that is important to us and that is true positive rate. TPR calculates the ratio of true positive voxels to the sum of true positive and false negative voxels. TPR is calculated using below equation:

$$TPR\ (M_R, M_A) = 2\frac{|M_R \cap M_A|}{|M_R \cap M_A| + |M_R \cap M_A^c|} \quad (2)$$

Our customized loss function can be written as follows:

$$Loss\ function = Weighted\ Dice\ (M_R, M_A) + TPR\ (M_R, M_A) \quad (3)$$

V. Data

In this study we use MS Lesion Challenge dataset [31], which is publicly available. This dataset is longitudinal which means it has several MRI scans for each patient at different time points. This dataset is also multi-modal. Image modality in MRI is a particular setting in the acquisition of the image which results in a particular image appearance. Different modalities have different contrasts. Each modality focuses on one aspect of MRI scans [32], making them to have complementary information. Because of that, most researchers design their networks to be multi-modalities. MRI scans in MS dataset, are 3D volumes with the shape of 181*217*181. Each subject has four different modalities including, flair, mprage, PD and T2. Below figure shows these modalities.

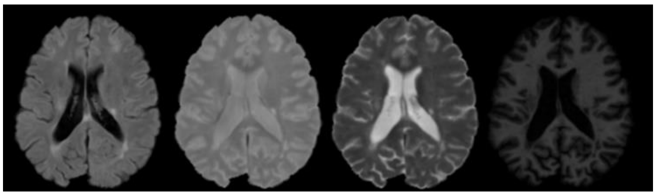

Fig. 7. Dataset MRI Modalities: Flair (Left), PD (Second Left), T2 (Second Right) and MPRAGE (Right).

The training dataset consists of MRIs from five patients, four of which have four time point and one of them have five time points. Test dataset has scans of 14 patients with overall 61 different subjects. Ground truth of the test dataset is not available and challenge participants can upload their segmentation masks onto the challenge website to calculate their proposed method performance.

VI. Metrics

In order to compare the results of our proposed method, we use dice coefficient overlap as our metric. Dice overlap is one the main metrics for comparison in medical domain, especially when the goal is predicting a binary mask. Dice overlap is the ratio of twice the number of overlapping voxels to the total number of voxels in each mask.

$$Dice\ (M_R, M_A) = 2\frac{|M_R \cap M_A|}{|M_R| + |M_A|} \qquad (4)$$

M_R is the ground truth mask and M_A is the mask generated by a particular algorithm. As you can see from the above equation, when M_R and M_A are identical, dice score will be one.

VII. Implementation

Since there is no available dataset with two different outputs, in this study we created our own proposed dataset. As mentioned before, we chose MS lesion challenge dataset, which contains brain MRIs along with lesion ground truth. In the second step, we provide brain tissue segmentation mask for each subject. We used Matlab's SPM tool [33] for this aim. SPM's brain tissue segmentation is based on image registration. Registration is the task of aligning two or more images [34]. In this tool, hundreds of brain images were aligned together to form a uniform brain scan. Then this referenced scan is used for tissue segmentation in new input MR images.

Fig. 8. Process of Brain Tissue Segmentation Using SPM Tool.

Figure 8 shows the process of brain tissue segmentation. SPM tool is designed for healthy brain scans, and since our original brain MRIs contain lesions, segmented brain structure is not accurate. To overcome this challenge, we proposed a strategy. In this strategy, we first train our network to predict lesions. Then we segment the scans into different tissues. Then we combine lesion parts with the white matter tissue to make sure that these lesions being segmented as white matter. Figure 9 shows this process.

As shown in figure 9, first step is to roughly segment the lesion parts. For this aim we used a regular 3D-UNET. We trained a patch-wised 3D-UNET, with the patch size of [64, 64, 64]. We used Adam [35] as our optimizer with the learning rate of 5e-4. We trained the network for 100 epochs, with a GTI 1050 Ti GPU. It is worth mentioning that we used a weighted dice loss function in this process.

For our multi-input and multi-output architectures, we also trained a patch-wised 3D-UNET with hyper-parameters that were mentioned above. The only difference apart from the architecture, is the loss function that we used for these networks.

VIII. Experiments and Results

In this section we report both quantitative and qualitative results of our proposed methods. To show that our method works better, we compare multi-task models with regular 3D-UNET.

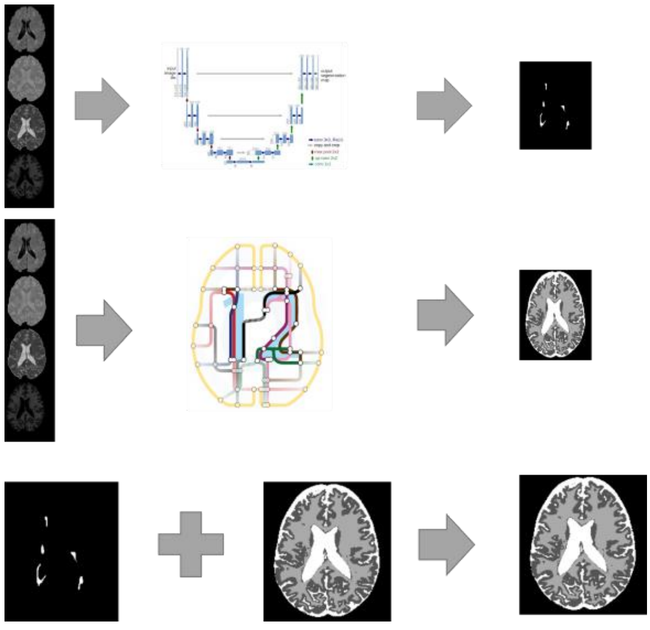

Fig. 9. Brain Tissue Segemntation in the Presence of Lesions.

A. Quantitative Results

To compare the results of our proposed models, in the quantitative part, we report the dice score of the predicted mask both on the train set and test set. In order to report the results on the train set, we used 5-fold cross-validation and the final dice score is the average between dice metric on each fold. Note that, in each fold, we used MRI scans of a single patient to avoid the network from seeing similar subjects in the training phase. We report the dice score on the test set as well. As mentioned before, the ground truth of the test set is not available, and to get the performance of your method, one needs to upload the predicted masks onto the challenge website. Table I shows the result of different methods on the train set.

TABLE I. AVERAGE DICE SCORE OF CROSS VALIDATION OF THREE DIFFERENT ARCHITECTURES ON TRAIN DATASET

	Deep-Learning Method		
	3D-UNET	***Multi-Input***	***Multi-Output***
Dice score	58.2	65.6	62

As shown in table I, both multi-input and multi-output, outperform the regular 3D-UNET model. You can also see that the performance of the multi-input model is higher than the multi-output network. We will explain these results in the discussion section.

Table II shows the performance of three architectures on the test set.

TABLE II. DICE SCORE OF THREE DIFFERENT ARCHITECTURES ON TEST DATASET

	Deep-Learning Method		
	3D-UNET	***Multi-Input***	***Multi-Output***
Dice score	49.7	54.8	52.1

Again like the train-set results, the best performance is achieved when we train the network in a multi-input manner.

B. Qualitative Results

In this section we show how different models, predict the segmentation mask. Figure 10 shows the prediction mask of different models on three different subjects.

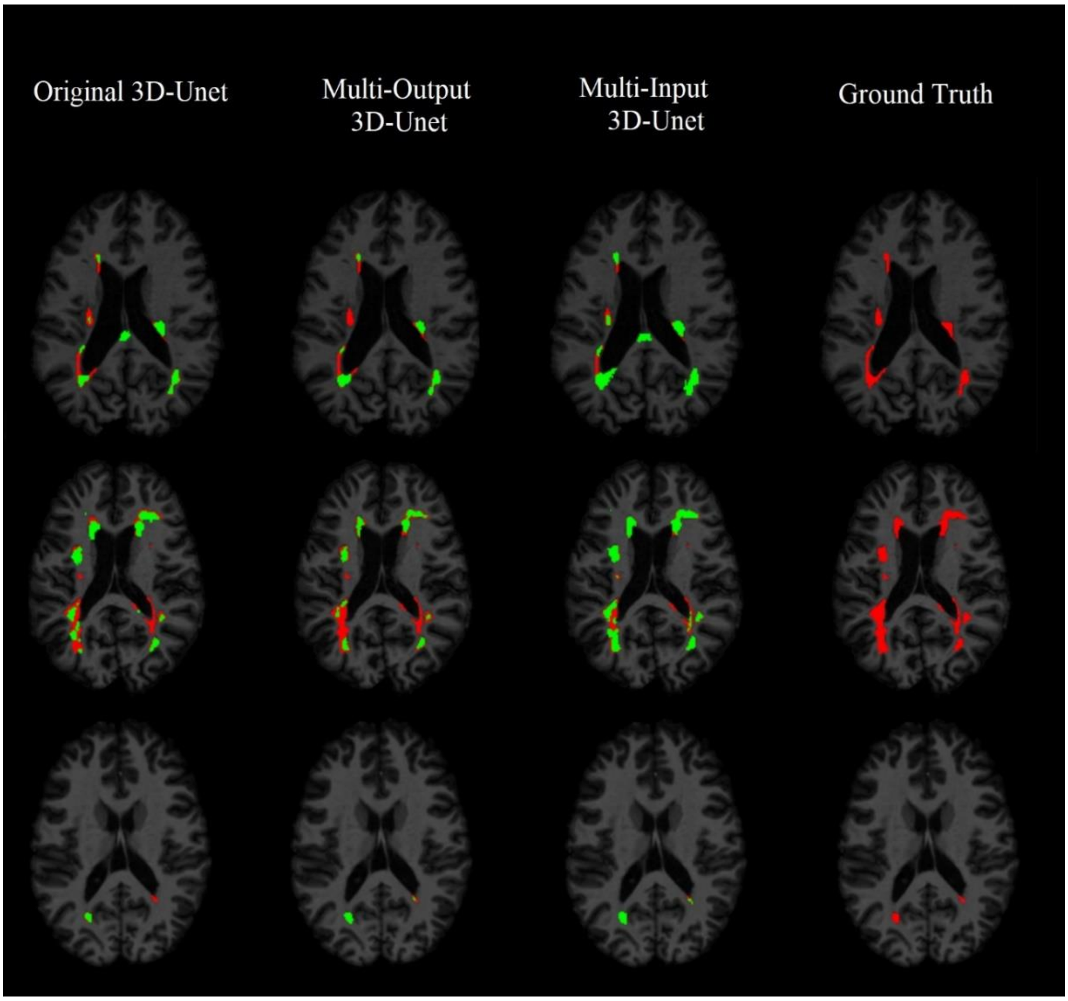

Fig. 10. Comparison of ground truth with output of different architectures. Red parts show the ground truth mask while green parts show the segmented area which is the output of each network. Each row is a different subject; Each column from left to right are: original 3D-Unet, Proposed Multi-Input 3D-Unet Architecture and Proposed Multi-Output 3D-Unet Architecture.

As you can see from figure 10, masks that were predicted by multi-task models, have lower false-positive voxels. This happens because the network implicitly learns that lesions only appear in the white matter. The picture also shows that the multi-input model is better at picking outline voxels. In other words, the multi-input model not only improves the false positive rate but also enhances the true positive rate.

IX. DISCUSSION AND CONCLUSION

In this study, we examined the power of auxiliary information to see whether it serves better as an input or output. Our findings show that even though multi-output architecture improves the performance, multi-input network, which uses the auxiliary information as an input to the network, achieves the best results. In this section, we explain why both multi-input and multi-output improve the performance. When we train our network in a multi-output manner, we are implicitly increasing the size of the dataset, which is proven to enhance the results. Another reason is since we are training the network regarding two loss functions, the network tries to learn a representation that suits both tasks well. In other words, features that can represent both tasks are chosen over features that are specific to only one task. This strategy assures that meaningful features have been selected so the algorithm generalizes better. In our case, when we train the network on both tasks, the network implicitly finds a mapping between white matter tissue and the location of lesions, which was more difficult for single-task models to understand.

After explaining why multi-output models perform better than single models, now we want to discuss the power of multi-input networks.

One of the reasons why multi-input performs better is that when we pass extra information to the network, the network uses this information as a fact but in the multi-output manner, not only the network tries to learn the relation between input and output, but also it should learn if there's any correlation between different outputs.

REFERENCES

[1] E. Krivov, M. Pisov, and M. Belyaev, "MRI augmentation via elastic registration for brain lesions segmentation," in International MICCAI Brainlesion Workshop. Springer, 2017, pp. 369–380.

[2] R. Merjulah and J. Chandra, "Segmentation technique for medical image processing: A survey," in 2017 International Conference on Inventive Computing and Informatics (ICICI). IEEE, 2017, pp. 1055–1061.

[3] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transforms for one-shot medical image segmentation," arXiv preprint arXiv: 1902.09383, 2019.

[4] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," Zeitschrift f¨ur Medizinische Physik, vol. 29, no. 2, pp. 102–127, 2019.

[5] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Mart´ı, and X. Llad´o, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," Artificial intelligence in medicine, 2018.

[6] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," arXiv preprint arXiv:1803.01164, 2018.

[7] Marleen De Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

[8] Rich Caruana. Multitask learning. Machine learning, 28(1):41–75, 1997.

[9] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning.In European conference on computer vision, pages 94–108. Springer, 2014.

[10] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre et al., "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," NeuroImage, vol. 148, pp. 77–102, 2017.

[11] Chelli N Devi, Anupama Chandrasekharan, VK Sundararaman, and Zachariah C Alex. Neonatal brain mri segmentation: A review.Computers in biology and medicine, 64:163–178, 2015.

[12] Sebastian Ruder. An overview of multi-task learning in deep neural networks.arXiv preprint arXiv:1706.05098, 2017.

[13] Isabelle Augenstein and Anders Søgaard. Multi-task learning of keyphrase boundary classification.arXiv preprintarXiv:1704.00514, 2017.

[14] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. InProceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 4822–4829, 2019.

[15] Mingsheng Long and Jianmin Wang. Learning multiple tasks with deep relationship networks.CoRR, abs/1506.02117,3, 2015.

[16] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogério Schmidt Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. InCVPR, volume 1, page 6, 2017.

[17] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3994–4003, 2016.

[18] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Learning what to share between loosely related tasks.arXiv preprint arXiv:1705.08142, 2017.

[19] Yu Zhang and Qiang Yang. A survey on multi-task learning.arXiv preprint arXiv:1707.08114, 2017.

[20] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever,and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. InInternationalConference on Medical Image Computing and Computer-Assisted Intervention, pages 478–486. Springer, 2016.

[21] Wufeng Xue, Andrea Lum, Ashley Mercado, Mark Landis, James Warrington, and Shuo Li. Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 276–284. Springer, 2017.

[22] Guosheng Hu, Li Liu, Yang Yuan, Zehao Yu, Yang Hua, Zhihong Zhang, Fumin Shen, Ling Shao, Timothy Hospedales, Neil Robertson, et al. Deep multi-task learning to recognise subtle facial expressions of mental states. InProceedings of the European Conference on Computer Vision (ECCV), pages 103–119, 2018.

[23] Xulei Yang, Zeng Zeng, Si Yong Yeo, Colin Tan, Hong Liang Tey, and Yi Su. A novel multi-task deep learning modelfor skin lesion segmentation and classification.arXiv preprint arXiv:1703.01025, 2017.

[24] Toan Duc Bui, Li Wang, Jian Chen, Weili Lin, Gang Li, and Dinggang Shen. Multi-task learning for neonatal brain segmentation using 3d dense-unet with dense attention guided by geodesic distance. InDomain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pages 243–251. Springer, 2019.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[26] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, volume 4, page 12, 2017.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.

[28] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep learning in medical image analysis and multimodal learning for clinical decision support, pages 240–248. Springer, 2017.

[29] Trygve Eftestøl. Controlling true positive rate in roc analysis. In2009 36th Annual Computers in Cardiology Conference(CinC), pages 353–356. IEEE, 2009.

[30] Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Co-maniciu, and Ghassan Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation.Computerized Medical Imaging and Graphics, 75:24–33, 2019

[31] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre et al., “Longitudinal multiple sclerosis lesion segmentation: resource and challenge,” NeuroImage, vol. 148, pp. 77–102, 2017.

[32] Hanzhang Lu, Lidia M Nagae-Poetscher, Xavier Golay, Doris Lin, Martin Pomper, and Peter CM Van Zijl. Routine clinical brain mri sequences for use at 3.0 tesla.Journal of Magnetic Resonance Imaging: An Official Journal of theInternational Society for Magnetic Resonance in Medicine, 22(1):13–22, 2005.

[33] John Ashburner, Gareth Barnes, C Chen, Jean Daunizeau, Guillaume Flandin, Karl Friston, Stefan Kiebel, JamesKilner, Vladimir Litvak, Rosalyn Moran, et al. Spm12 manual.Wellcome Trust Centre for Neuroimaging, London, UK, page 2464, 2014.

[34] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier et al., “Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration,” Neuroimage, vol. 46, no. 3, pp. 786–802, 2009.

[35] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.