Unified Representation Network with Missing Modalities Analysis

Kadir Gökdeniz, Mehmet Bayram Alpay
Department of Computer Engineering
Ankara University
Email: 20290344@ankara.ogrenci.edu.tr
20290310@ankara.ogrenci.edu.tr

I. Problem definition

Methods used for diagnosing clinical diseases are of vital importance. Segmenting images is a routine practice in diagnosing clinical diseases. However, due to the different approaches and opinions of experts during diagnosis, diseases can be interpreted in various ways [1]. The research titled ”A Unified Representation Network for Segmentation with Missing Modalities” focuses on the segmentation analysis of medical images, for brain tumor segmentation, aiming to reduce errors that may arise from these different approaches [2].

The modality dropout method is commonly used in medical image segmentation [3]. The reason for using missing modalities in the examination of medical images is that different modalities have different effects on the segmentation process, which can lead to over-reliance on a single factor and reduce the accuracy of the results. Modality dropout aims to reduce this reliance. This method balances the segmentation process by randomly removing modalities from the image. It should also be noted that in daily life, we often attempt to solve segmentation problems by assuming that all modalities are present in photographs. However, this assumption may not always hold true in real-life scenarios. The goal is to ensure that the performance remains sufficiently good even in situations where this assumption does not hold.

The modality dropout method can cause the input and output data to become highly independent from each other by leading to the loss of modalities. As an alternative solution to this problem, the research introduces the unified representation network. The unified representation network transforms the data into a common representation despite the missing modalities and synthesizes images by ensuring their proper mapping. It is noted to produce better results compared to using only modality dropout.

II. Comparison

To present the advantages of the main research article, we will compare it with methods used in two different literature studies.

One significant advantage of the Unified Representation Network is its ability to train using multiple datasets. On the other hand, pretrained models may have the potential to complete learning from data more sensitively. As detailed in the results section, pretrained models trained only on BRaTS and BRaTS-HCP datasets were separately treated in this study, and their results were listed accordingly. Success in image synthesis significantly varied depending on the headings.This feature varies compared to many studies in the literature and allows for the enumeration of the advantages and disadvantages of providing pretraining with different datasets [4].

Although the method section will be discussed in detail, it is worth mentioning the structure of the unified representation network here. One additional advantage provided by the Unified Representation Network is its simple structural design. Therefore, due to the nature of the Unified Representation Network, it has the ability to perform segmentation in a short period.

On the other hand, in the study [5], feature generation generator and correlation constraining blocks are added to the structure of the Unified Representation Network. Figure 1 visualizes the flow of the other study. In brief, the feature-enhanced generator creates a new modality using the modalities present in the photograph. The generated modality and the other modalities in the photograph are sent to the correlation-constraining block. The correlation-constraining block attempts to perform segmentation by using the modalities, incorporating missing modalities into consideration. The results of this research indicate that the obtained dice scores are better than the main research results. Since the feature-enhanced generator and correlation-constraining blocks add a more complex structure to the Unified Representation Network, the segmentation process takes longer.

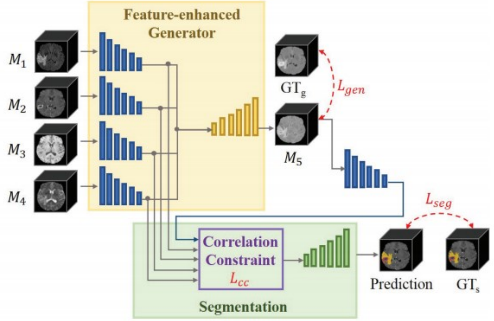


Fig. 1. The flow diagram of the research [5].

Study [6] proposed a new model for brain tumor segmentation. This model, unlike the Unified Representation Network (URN), has a multi-encoder structure that allows learning separate feature representations from each imaging modality. Additionally, similar to the structure in study [5], it uses a correlation model to learn relationships between different modalities. The attention-based fusion block combines feature representations using channel attention and spatial attention, making the feature representations more selective. Moreover, it utilizes reconstruction decoders to enhance segmentation performance. This study successfully coped with missing modalities and provided more effective learning with the correlation feature. The attention-based fusion blocks improved segmentation performance by behaving more selectively in feature representations. All of these features contributed to achieving better results than study [2].

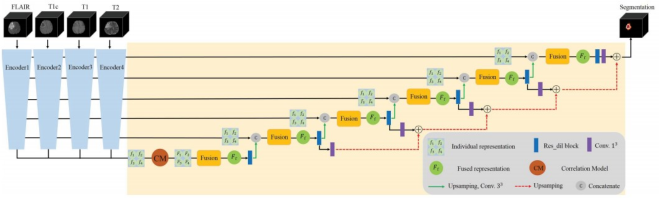


Fig. 2. The flow of model [6].

III. Input

In the research [2], two different datasets were utilized to take advantage of the ability of the unified representation network to leverage multiple training datasets. Firstly, a dataset used in the 2018 Multimodal Brain Tumor Segmentation (BRATS) competition was selected. This dataset contained pre-trained MR images from 285 glioblastoma patients. The second dataset comprised MR images and was obtained from 1108 healthy individuals. This dataset was sourced from the Human Connectome Project (HCP).

For tumor segmentation, the dataset was divided into three classes, commonly used in other studies as well: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). The WT class also included peritumoral edema data. Similar to other MR imaging studies, both datasets included FLAIR, T1, T1c, and T2 modalities. The modalities were randomly left out in the unified representation during training. 70% of the data was used for training, while the remaining 30% was reserved for testing.

T1, T1c, T2 and Flair were the modalities we used in our project. These modalities are different methods commonly used in magnetic resonance imaging techniques. Each one shows different tissue and different regions.

T1: It is the modality that shows the anatomical structure and fat tissue well. It helps us understand the gray-white matter distinction of the brain. In this modality, fatty tissues and protein-rich parts appear brighter.

T1c: This modality shows tumor tissues and areas with abnormal blood flow brighter. It makes pathological areas more visible and is used to determine the location of tumors.

T2: It is an ideal modality to show fluid-containing structures and edema. Fluid-filled spaces appear more clearly. Solid tissues appear dark and it becomes easier to distinguish.

Flair: It is used to show lesions and pathological conditions in the brain. It is a modality that helps detect small lesions by pressing on fluid-filled areas. Cerebrospinal fluid appears dark, while edema and lesions appear brighter.

These four modalities are critical in the diagnosis and treatment of brain tumors and other neurological conditions. Each modality shows different pathological conditions and anatomical details, which helps doctors make an accurate diagnosis. When we evaluate it in terms of our own project and the results we get, the Flair modality seems to have a much greater effect on the results we get compared to other modalities. The results we obtained in experiments involving the Flair modality show more positive figures.

IV. Method

The method part mainly focus on 3 sub-parts:

A. Baseline

This network architecture typically found in convolutional neural networks begins by passing through a 3x3 convolution, definition is obtained for architecture by research [7]. Since these convolutions are without padding, the width and height of the image decrease by two after each convolutional step. Subsequently, it undergoes a 2x2 max pooling operation with a stride of 2. During max pooling, the number of feature channels is doubled.

Additionally, the network passes through upper convolutional layers. These layers consist of a 2x2 convolutional layer that reduces the number of feature channels by half and is merged with the cropped feature map. In addition to upper convolutions, it passes through layers consisting of two 3x3 convolutions. When reaching the final layer, a 1x1 convolution is used to match the 64-component feature vector to the desired number of classes. The architecture is completed with a total of 23 convolutional layers.

In the mentioned study, the leakyReLU activation function is used for each pooling operation. Additionally, a structure containing one convolutional layer for each resolution level is preferred. Batch normalization is commonly used in many problems to improve deep learning performance. This feature has also been utilized in the [2].

During the training phase, the Caffe implementation employs stochastic gradient descent (SGD). Since the number of input data exceeds the number of output data, processing loads may vary across convolutional layers, particularly. To balance this effect and optimize GPU usage, large batch sizes are typically used. Additionally, current optimization processes are updated using high momentum coefficients (0.99).

At the baseline, the energy function is calculated as a combination of the softmax function applied to the final feature map and the pixel-wise cross-entropy loss function. The softmax function is determined by ak(x), representing the activation in feature channel k at each pixel position. Here, K denotes the number of classes, and pk(x) approximates the maximum function. Subsequently, the cross-entropy penalizes deviations of p(x)(x) from 1 at each position.

Additionally, a weight map is computed during training to assign higher importance to certain pixels. The separation boundary is computed using morphological operations, and then the weight map is derived using wc to balance class frequencies and d1 and d2 to represent distances to the nearest cell borders. In a deep network, having well-initialized weights is crucial to prevent certain parts from dominating with excessive activations while others contribute minimally. Ideally, initial weights should be adjusted to ensure that each feature map in the network has approximately unit variance.

For our architecture (comprising alternating convolution and leakyReLU layers), initial weights can be drawn from a Gaussian distribution with a standard deviation of p2/N, where N represents the number of incoming nodes of one neuron. For example, for a previous layer with a 3x3 convolution and 64 feature channels, N = 9 * 64 = 576.

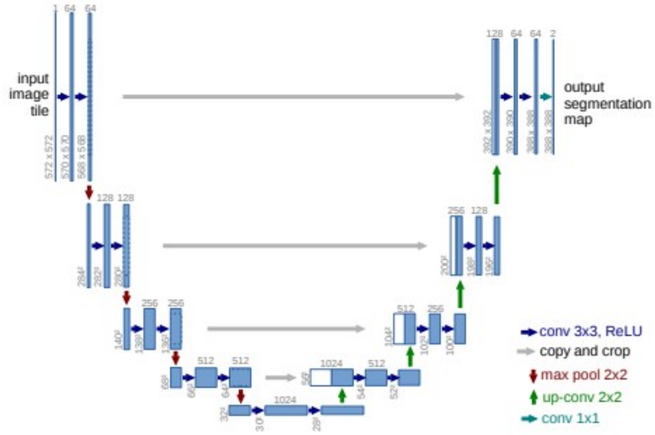


Fig. 3. The architecture of study [7].

B. Modality Dropout

Modality dropout is a method used during medical image synthesis. The modality dropout method randomly zeros out input channels. Models equipped with this method are compelled to learn with missing modalities. One of the most significant advantages of modality dropout is its ability to predict inconsistencies between training and test data. This method allows segmentation of modalities independently of the relationships between different modalities.

Modality dropout is implemented as follows: Let's assume that k modalities will be dropped. We aim for an exponential decrease in the number of dropped modalities. The probability of modalities dropping is denoted as pθ(k), where θ is a constant between 0 and 1. Due to the tendency of the normalization factor to have a geometric distribution, we obtain equation 1.

$$p_\theta(k) = \frac{(1-\theta)\theta^k}{1-\theta^{N+1}}, \quad k = 0, \ldots, N_{\max}.$$

Fig. 4. Equation 1 [2].

C. Unified Representation Network

An important feature of the Unified Representation Network is its ability to learn from a variable number of input values. Traditional approaches typically struggle to accommodate variable data inputs. The Unified Representation Network consists of three components: encoding, fusion, and decoding.

Encoding - The structure used is essentially the model introduced in the baseline section. The unified representation shares the same width and height as the input. We used 16 channels. Fixed parameters were employed to standardize the output channels of each encoder. This modification is a crucial factor for performance, simply normalizing the data to have a mean of zero and a variance of one.

Fusion - The fusion module may struggle to perform mapping with considering the dropped modalities due to modality dropout. Additionally, the fusion module needs to adapt to a variable number of input data. During fusion, it is desired that the results are intensive. For this operation to be successful, the fusion function must be mathematically invertible. In this study, the function f(x) = x is used for simplicity.

The fusion module requires regularization to perform its operations effectively. In cases where all modalities are dropped, the fusion module may fail to produce the desired outcome. Therefore, it is necessary to include the variance in the loss function to ensure that there are always at least two modalities available.

Decoding - The decoding module synthesizes images based on the unified representation network. Decoding blocks have the advantage of imposing minimal restrictions on the data, which significantly increases the amount of usable data. The training are completed by using different decoding structures. Decoding blocks consist of two residual blocks and a final 1x1 convolutional layer.

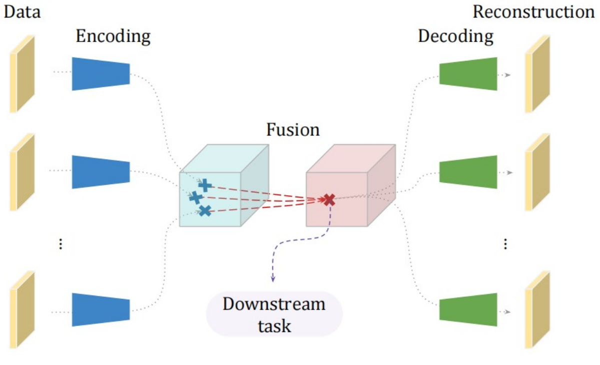
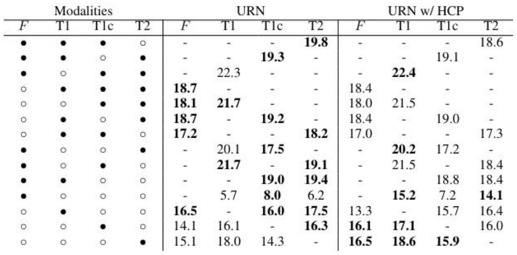


Fig. 5. Unified Representation Network [2].

V. Results

The results were evaluated in 3 different categories as whole tumor, enhancing tumor and tumor core areas. Evaluation results were measured by looking at the segmentation results of the models. The results of different modalities evaluated with a total of 4 different models, which started with baseline models and then were produced using the unified representation network structure, are given in the Table 1. The metric on which the results are based here is called Dice scores, and it can be said that the model with a higher Dice score performs better. An inference can be made that the model with a better Dice score also has a better accuracy score, and in this way, it becomes easier to explain and interpret the experimental results.

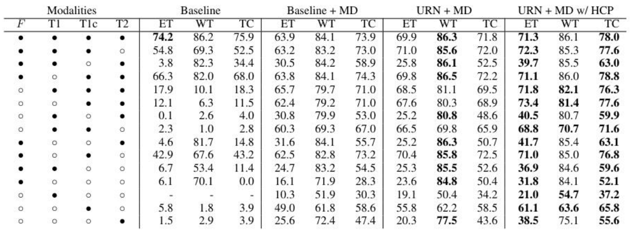
Table 2: Peak signal-to-noise ratio of synthesized BRATS images.



Modalities				URN				URN w/ HCP			
F	T1	T1c	T2	*F*	T1	T1c	T2	*F*	T1	T1c	T2
●	●	●	○	-	-	-	**19.8**	-	-	-	18.6
●	●	○	●	-	-	**19.3**	-	-	-	19.1	-
●	○	●	●	-	22.3	-	-	-	**22.4**	-	-
○	●	●	●	**18.7**	-	-	-	18.4	-	-	-
○	○	●	●	**18.1**	**21.7**	-	-	18.0	21.5	-	-
○	●	○	●	**18.7**	-	**19.2**	-	18.4	-	19.0	-
○	●	●	○	**17.2**	-	-	**18.2**	17.0	-	-	17.3
●	○	○	●	-	20.1	**17.5**	-	-	**20.2**	17.2	-
●	○	●	○	-	**21.7**	-	**19.1**	-	21.5	-	18.4
●	●	○	○	-	-	**19.0**	**19.4**	-	-	18.8	18.4
●	○	○	○	-	5.7	**8.0**	6.2	-	**15.2**	7.2	**14.1**
○	●	○	○	**16.5**	-	**16.0**	**17.5**	13.3	-	15.7	16.4
○	○	●	○	14.1	16.1	-	**16.3**	**16.1**	**17.1**	-	16.0
○	○	○	●	15.1	18.0	14.3	-	**16.5**	**18.6**	**15.9**	-

Modality dropout was used when training the model, and this increased the performance evaluated in the test examples, so it came to the fore as one of the recommended methods. While training the models presented, there were various benefits of training on multiple datasets instead of using only a single dataset. These are having a stronger training space and the possibility of creating stronger visual structures by synthesizing images. We call the images synthesized with different images from different datasets reconstructed images, and we see the results in Table 2.

Table 1: Dice scores for different models on BRATS.



Modalities				Baseline			Baseline + MD			URN + MD			URN + MD w/ HCP		
F	T1	T1c	T2	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
●	●	●	●	**74.2**	86.2	75.9	63.9	84.1	73.9	69.9	**86.3**	71.8	**71.3**	86.1	**78.0**
●	●	●	○	54.8	69.3	52.5	63.2	83.2	73.0	71.0	**85.6**	72.0	**72.3**	85.3	**77.6**
●	●	○	●	3.8	82.3	34.4	30.5	84.2	58.9	25.8	**86.1**	52.5	**39.7**	85.5	**63.0**
●	○	●	●	66.3	82.0	68.0	63.8	84.1	74.3	69.8	**86.5**	72.2	**71.1**	86.0	**78.8**
○	●	●	●	17.9	10.1	18.3	65.7	79.7	71.0	68.5	81.1	69.5	**71.8**	**82.1**	**76.3**
○	○	●	●	12.1	6.3	11.5	62.4	79.2	71.0	67.6	80.3	68.9	**73.4**	**81.4**	**77.6**
○	●	○	●	0.1	2.6	4.0	30.8	79.9	53.0	25.2	**80.8**	48.6	**40.5**	80.7	**59.9**
○	●	●	○	2.3	1.0	2.8	60.3	69.3	67.0	66.5	69.8	65.9	**68.8**	**70.7**	**71.6**
●	○	○	●	4.6	81.7	14.8	31.6	84.1	55.7	25.2	**86.3**	50.7	**41.7**	85.4	**63.1**
●	○	●	○	42.9	67.6	43.2	62.5	82.8	73.2	70.4	**85.8**	72.5	**71.0**	85.0	**76.8**
●	●	○	○	6.7	53.4	11.4	24.7	83.2	54.5	25.3	**85.5**	52.6	**36.9**	84.6	**59.6**
●	○	○	○	6.1	70.1	0.0	16.1	71.9	28.3	23.6	**84.8**	50.4	**31.8**	84.1	**52.1**
○	●	○	○	-	-	-	10.3	51.9	30.3	19.1	50.4	34.2	**21.0**	**54.7**	**37.2**
○	○	●	○	5.8	1.8	3.9	49.0	61.8	58.6	55.8	62.2	58.5	**61.1**	**63.6**	**65.8**
○	○	○	●	1.5	2.9	3.9	25.6	72.4	47.4	20.3	**77.5**	43.6	**38.5**	75.1	**55.6**

Some of the important building blocks of the project were as follows. By adding modality dropout, better performance results emerge when faced with missing modality, because dropout loosens the strong ties between multimodals, gives importance to the individual properties of modals, and allows creating a more stable model against this missing modality problem. Another is that some modalities are more informative than others. For example, it is clear from the results that FLAIR is of the highest importance for all experiments, so we can call FLAIR the modality of high importance.

VI. Advantages

One of the most important advantages of the neural networks used is that they are sensitive to different inputs, because the models may be helpless against a variable number of input modalities and it would be necessary to train a different neural network against each possible combination of inputs, and compare the results.

Developing neural networks that solve this problem provides us with a sidestep contribution to the combinatorial explosion.

The main contribution of the project is to provide solutions to frequently encountered problems such as missing modality rather than offering different network architectures. By using the standard U-net and its environments, the advantage of providing robustness against missing modalities and mapping each input channel of the U-net block to a unified representation has been discovered.

It has been understood that removing random missing modalities makes the training process more effective and increases the synthesis between these MR images.

Neural networks with trainable encoding and decoding structures for unsupervised learning problems have the advantage of benefiting from various and multiple sources. It has enabled resources such as text and images to simultaneously contribute to the educational space.

There is a specific encoding structure for each modality, and the results from this function to standardize all outputs using batch-normalization. Regardless of the number of modalities, it brings different modalities into a dense feature context in the fusion part, and the downstream task takes place here. Decoding, on the other hand, performs task-specific operations based on unified representation. This structure allows us to query missing modalities in 4 different models such as end-to-end baseline, baseline with modality dropout, URN with modality dropout and pre-trained URN with modality dropout.

VII. Disadvantages and improvement

The modality dropout method may have some disadvantages, for example, since it randomly resets the input channels, this may not fully detect inconsistencies in real-world data in some cases. It may also ignore the relationships between missing modalities and cause this model to produce misleading results.

The disadvantage of Unified Representation Network is that it has a more complex structure compared to modality dropout and requires a longer time for training, which means increases its cost. Additionally, as the number of missing modalities increases, the risk of synthesizing the data increases and the model begins to fail to produce the desired result in the fusion module.

Medical imaging data is very difficult to collect and process, and it is very costly to bring together data from different sources and involving different modalities. Resolving inconsistencies in data coming from different sources and inconsistencies reflected in the results is a significant disadvantage for this problem.

In the study [5], the use of feature generation generator and correlation constraining blocks helps to improve the dice score performance. New features can be generated and learned through the use of two encoding and decoding stages. In study [6], a correlation model is used to learn from different modalities. In addition, attention-based fusion blocks make feature representations more selective. Segmentation performance is enhanced with the help of reconstruction decoders. All these features contribute to achieving better results compared to study [2].

The results helped us understand the problem and performance, for example Dice scores contributed significantly to our performance comparison, but if metrics such as sensitivity, specificity or AUC score were used in addition to Dice scores, this would be more effective in clinical environments where false positives and false negatives are important. It provides a detailed perspective. The training process creates difficulties because it takes a long time. Techniques such as knowledge distillation or model compression without sacrificing performance may be a good option. Additionally, distributed systems and hardware accelerators can be solutions that can reduce the time for more frequent experiments. It may be useful to use transfer learning techniques, for example, a pre-trained model trained with large brain segmentation tumor data can contribute to better performance with the data we have.

References

[1] G. Sethuram Rao and D. Vydeki, "Brain Tumor Detection Approaches: A Review," 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 479-488, *[Online].* Available: doi: 10.1109/ICSSIT.2018.8748692.

[2] K. Lau, J. Adler, and J. Sjölund, "A unified representation network for segmentation with missing modalities," arXiv preprint arXiv:1908.06683, 2019.

[3] Brian E. Perron, Charlotte L. Bright,The influence of legal coercion on dropout from substance abuse treatment: Results from a national survey, Drug and Alcohol Dependence, Volume 92, Issues 1–3, 2008, Pages 123-131, ISSN 0376-8716,*[Online].* Available: https://doi.org/10.1016/j.drugalcdep.2007.07.011.

[4] Özgenel, Ç. F., and A. Gönenç Sorguç. "Performance comparison of pretrained convolutional neural networks on crack detection in buildings." Isarc. proceedings of the international symposium on automation and robotics in construction. Vol. 35. IAARC Publications, 2018.

[5] Tongxue Zhou, Stéphane Canu, Pierre Vera, Su Ruan,Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities,Neurocomputing,Volume 466,2021,Pages 102-112,ISSN 0925-2312,*[Online].* Available: https://doi.org/10.1016/j.neucom.2021.09.032.

[6] T. Zhou, S. Canu, P. Vera and S. Ruan, "Latent Correlation Representation Learning for Brain Tumor Segmentation With Missing MRI Modalities," in IEEE Transactions on Image Processing, vol. 30, pp. 4263-4274, 2021,*[Online].* Available: doi: 10.1109/TIP.2021.3070752.

[7] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)