2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

WEAKLY-SUPERVISED BRAIN TUMOR CLASSIFICATION WITH GLOBAL DIAGNOSIS LABEL

Yufan Zhou ⋆ *Zheshuo Li* ⋆ *Chunwei Ma* ⋆
Changyou Chen ⋆ Mingchen Gao ⋆ Hong Zhu † Jinhui Xu ⋆

⋆ Department of Computer Science and Engineering, University at Buffalo, SUNY
† School of Medical Information, Xuzhou Medical University

ABSTRACT

There is an increasing need for efficient and automatic evaluation of brain tumors on magnetic resonance images (MRI). Most of the previous works focus on segmentation, registration, and growth modeling of the most common primary brain tumor gliomas, or the classification of up to three types of brain tumors. In this work, we extend the study to eight types of brain tumors where only global diagnosis labels are given but not the slice-level labels. We propose a weakly supervised method and demonstrate that inferring disease types at the slice-level would help the global label prediction. We also provide an algorithm for feature extraction via randomly choosing connection paths through class-specific autoencoders with dropout to accommodate the small-dataset problem. Experimental results on both public and proprietary datasets are compared to the baseline methods. The classification with the weakly supervised setting on the proprietary data, consisting of 295 patients with eight different tumor types, shows close results to the upper bound in the supervised learning setting.

Index Terms— Autoencoder, small dataset, feature ensemble

1. INTRODUCTION

Brain tumors are among the most fatal cancers. Around 25 per 100,000 adults are diagnosed with primary tumors of the brain or nervous system and approximately one-third of the tumors being malignant [1]. Many different types of brain tumors exist, such as gliomas, pituitary and meningiomas. Magnetic Resonance Images (MRI) is a clinical routine frequently used for brain tumor detection and classification.

While deep learning based algorithms have achieved enormous success in medical image analysis fields, they usually heavily rely on fully annotated data. Annotating large amount of medical data usually requires expert domain knowledge, and is tedious, time-consuming and not realistic to obtain. In this paper, we focus on overcoming the data hungry problem where only limited training data are available or only global diagnosis is available but costly detailed annotation is not available. The global diagnosis information is less informative than detailed annotation mask, but can be retrospectively obtained with significantly lower cost.

Analyzing brain tumor on MR images is an important topic in medical imaging, which has motivated advanced deep learning techniques for classification, detection, segmentation, registration, retrieval, image generation and enhancement [2]. Many research efforts have been devoted to brain tumor segmentation [3, 4, 5]. Another track of work is to classify an MR volume directly or classify MR slices given corresponding supervised labels [6, 7, 8].

Most existing works mainly focus on supervised learning in brain tumor classification. Specifically, given global diagnosis labels, a volumetric MRI is considered as a single data point [8], whereas other works [6, 7] consider each MRI slice as a single data point, which requires labels to each slice. Our work focuses on predicting slice labels using only the given global labels of the training data. The global labels are then inferred by combining the predicted slice labels. This not only makes the dataset labeling free, but also significantly reduces the chance of overfitting, an issue often arising in treating a whole MRI volume as a single data point. The idea of eliminating the human labeling effort has attracted a considerable amount of attention in the research community in brain tumor segmentation, histopathology image classification and natural image segmentation [9, 10, 11].

We propose a weakly supervised learning method to overcome the challenge of acquiring expensive slice-level labels. Weakly supervised learning represents the training scheme where only the diagnosis for a whole MRI volume is given, but the detailed labels for each slice are unknown during training. We treat each MRI slice as one data point, whose label could be healthy or tumor. The problem is then transformed to an optimization problem for deciding the label of each slice between the global diagnosis label and the healthy label.

Instead of using a traditional convolutional neural network (CNN) performing feature learning and classification

Yufan Zhou and Zheshuo Li are equally contributed co-first authors; Hong Zhu and Jinhui Xu are co-corresponding authors. This work was supported in part by NSF through grants CCF-1716400 and IIS-1910492.

simultaneously, we propose a hierarchical model to learn features using class-specific autoencoders separately from the classification network. Inspired by Dropout [12] and DropConnect [13], we propose Block-connection-dropout for regularizing the process of feature extraction. These ideas are intended to perform implicit data augmentation, introduce noises and regularize the model, such that they can improve the generalization ability of the networks, and thus give improved testing performance.

Our contributions can be summarized as follows:

- We propose a weakly supervised learning method that treats each MRI slice as a data point and infers the patient diagnosis label by aggregating the slice label prediction.
- We propose a feature extraction scheme called Multi-RAED. It is trained with class-specific autoencoders, and extracts features by selecting random paths within those autoencoders using block-connection-dropout.
- We have collected a dataset with 295 patients of eight types of tumors. The proposed algorithms are demonstrated to be effective compared to fully supervised learning.

2. METHODS

2.1. Weakly-supervised Learning

Different from existing public datasets with detailed slice-level labels [6], our dataset has only global labels in patient level. We propose to use a weakly-supervised learning method to locate the slices where the tumor presents, and subsequently to help predict the patient-level labels. Our method is based on the observation that for a type y MR volume, its slices can be labeled as either y or healthy. We augment the label space by adding a label representing a healthy MRI slice, such that the label of each slice can be represented by a one-hot vector of length $K+1$, where K denotes the total number of diseases in the dataset.

Given an MR image $\mathbf{x}$, we let $(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)})$ denote its M slices. The corresponding ground truth patient label is represented by a one-hot vector $\mathbf{t}$. Its predictions are denoted as $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}^{(1)}, \cdots, \tilde{\mathbf{y}}^{(M)})$, where $\tilde{y}^{(i)}$ is the prediction of the i-th slice which is also a one-hot vector. Let $\bar{y} = \frac{1}{M}\sum_{i=1}^{M} \tilde{y}^{(i)}$ be the average prediction through all slices. We model the weakly supervised learning problem as an optimization problem with the following objective function for an MR image:

$$\mathcal{L}(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{t}) \triangleq -\sum_{j=1}^{K} t_j \log \bar{y}_j + (\alpha \|\bar{y}_{1:K}\|^2 + \beta \bar{y}_{K+1}^2), \quad (1)$$

where K is the number of classes. The first term in the equation $-\sum_{i=j}^{K} t_j \log \bar{y}_j$ is the cross entropy loss between $\mathbf{t}$ and $\bar{\mathbf{y}}_{1:K}$, which emphasizes that among all the predicted $\tilde{y}^{(i)}$'s, if they are not predicted as healthy slices, they are constrained to be the same as the ground truth tumor type. By setting $\alpha > \beta$, the second term avoids the scenario where all the slices are labeled as the ground truth tumor type. This is because if so, the second term reduces to αM. But if one slice is labeled as healthy (*i.e.,* class $K+1$), the second term becomes $\alpha(M-1)+\beta < \alpha M$, leading to a smaller loss. Thus, the objective function ensures that some slices are predicted as the ground true tumor type, but not all of them are expected to be categorized into the same class.

2.1.1. *Test-time MR volume classification.*

Our proposed model is able to predict slice labels without slice training labels. The prediction of an MR volume, denoted as $\bar{\mathbf{y}}$, will be the class associated with the largest $\bar{y}_i$, *i.e.*, $k = \arg\max_i \bar{y}_i$.

2.2. Feature Extraction via Autoencoder-Dropout

We propose to train two networks for the feature extraction and classification respectively, as shown in Fig.1. An autoencoder is first trained to extract features from the training data, and then a classifier is built based on the extracted features. Because the extracted features are set to endow smaller dimensions than the original data, the classifier is expected to be trained faster without too much performance loss.

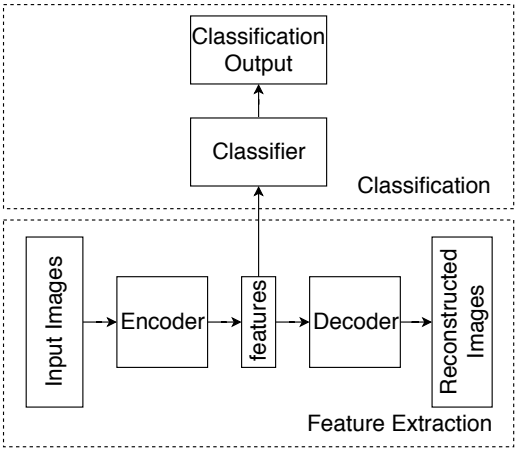

Fig. 1. Autoencoder-classification model

2.2.1. *Multi-Autoencoders (MultiAE)*

Instead of using a single autoencoder model for all classes, we propose to use multiple class-specific autoencoders to learn discriminative features, each corresponding to one class. We denote this model as MultiAE, and the features extracted from one autoencoder is called a *feature group*. Once all the autoencoders are trained, features from all autoencoders can

be used as features for classification. By training on class-specific data, the MultiAE is able to encode class information in latent features, potentially leading to better discrimination ability. This can also be considered as training data augmentation to alleviate the problem of limited training data. We use dense blocks to build an autoencoder, to preserve information from both low-level features and high-level features.

2.2.2. Random combination of autoencoders with dropout (MultiRAED)

Inspired by the feature learning in CNN, we design a similar hierarchical feature representation learning paradigm. In CNN, features are learned hierarchically by combining low-level features such as edges and corners to high-level features such as complicated shapes or objects. In our class-specific autoencoders, we randomly pick paths of the dense blocks to learn features. If we have m autoencoders with each having n dense blocks, there are m^n possible combinations for constructing the hierarchical feature representations. An example of two autoencoders with two dense blocks each is illustrated in Fig. 2, where four feature groups are generated to fit into the classifier. Each input data sample can be augmented to m^n feature groups.

The number of feature groups will increase exponentially, leading to redundant input for the classifier. To reduce the complexity of the extracted features and also prevent their co-adaptation, we propose a block-connection-dropout architecture inspired by DropConnect [13]. When extracting features to train the classifier, we randomly drop some possible connections between the blocks. This model is denoted as **MultiRAED** in later content. Specifically, denote the mapping by the classifier implemented by a neural network as $\mathbf{y} = f(\mathbf{x})$ with input features $\mathbf{x}$. The output $\mathbf{y}$ is an unnormalized vector of length K, where K represents the number of classes. When the input contains only one feature group $\mathbf{x}$, the probability of the input belonging to class i is calculated as: $P_i = \frac{e^{y_i}}{\Sigma_{j=1}^{K} e^{y_j}} = \frac{e^{[f(\mathbf{x})]_i}}{\Sigma_{j=1}^{K} e^{[f(\mathbf{x})]_j}}$, where y_i represents the i^{th} element in the vector $\mathbf{y}$. When p feature groups were selected after dropout some feature groups, denoted as $\mathbf{x^1}$ ... $\mathbf{x^p}$, the probability of the input belonging to class i is:

$$P_i = \frac{e^{\Sigma_{t=1}^{p} y_i^t}}{\Sigma_{j=1}^{K} e^{\Sigma_{t=1}^{p} y_j^t}} = \frac{e^{\Sigma_{t=1}^{p} [f(\mathbf{x^t})]_i}}{\Sigma_{j=1}^{K} e^{\Sigma_{t=1}^{p} [f(\mathbf{x^t})]_j}} \ .$$

3. EXPERIMENTS

To test our method, we have created so far the largest brain tumor dataset in terms of number of tumor types, which contains 8 tumor types with 295 patients. There are 50 Glioma, 50 Meningioma, 44 Metastases, 26 Lymphoma, 35 Prolactinoma, 22 Ependymoma, 22 Medulloblastoma, and 46 Acoustic Neuroma patients in the dataset. We adopt a 7-fold cross validation setting to estimate the performance, *i.e.*, the dataset is partitioned into a training set (72%), a validation set (14%) and a test set (14%) in every training round. We also test our method on a public dataset [6]. The public dataset contains 3 tumor types and about 200 patients.

Table 1. Classification accuracy on public dataset

2D-SingleAE	2D-MultiAE	2D-MultiRAED	[6]	[7]
89.62% ±3.22%	90.87% ±1.63%	**91.80%** ±**2.80%**	91.28%	86.56%

3.1. Parameter Settings

Every autoencoder consists of 3 dense blocks in encoder and decoder part respectively, and the classifier consists of 2 dense blocks. All the dense blocks have 6 convolutional layers inside. To prevent overfitting, dropout with dropout rate 0.1 and weight decay with hyper-parameter $1e-4$ are implemented. Learning rate decay is also used to improve the performance, learning rate will multiply 0.3 when 50% of the training process and 75% of the training process is finished respectively. We randomly select 27 feature groups (out of $8^3 = 512$ possible feature groups) for each MR volume due to the memory limitation. α and β in Eq.(1) are set to be 1 and 0.01 respectively. Our implementation is based on Tensorflow and Nvidia Titan Xp GPU.

3.2. Experimental Results and Discussion

We first test some of our models on the public dataset [6]. Because the public dataset only contains slice-level data, the weakly-supervised method is not applicable, we only test Multi-AE and Multi-RAED. For comparison, we also test the result using single autoencoder (denoted as Single-AE). The results are shown in Table 1. The experiments on the public dataset shows that the separation of feature learning and classification will achieve similar accuracy as the direct classification.

To demonstrate the effectiveness of the weakly supervised learning and the MultiRAED models, we test five different models on two tasks of our collected dataset: 3-type and 8-type brain tumor classifications. For the 3-tumor-type case, Glioma, Meningioma, and Metastases patients are selected to be consistent with the public dataset. In both cases, five models are tested in the following experiments:

- **DenseNet**: This model is served as a baseline for classifying 2D MR slices or 3D MR volumes directly using a DenseNet structure [14].
- **MultiRAED**: This model classifies 2D MR slices or 3D MR volumes with MultiRAED features extraction. Weakly supervised learning is not used in the first two models.

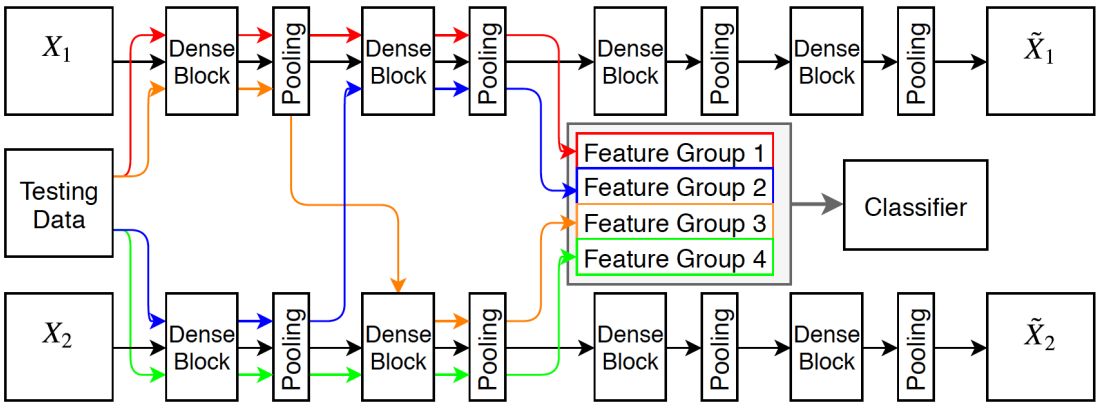

Fig. 2. Illustration of proposed model with two autoencoders and two classes of data. X_1 and X_2 denote training data from two different classes. $\tilde{X_1}$ and $\tilde{X_2}$ are corresponding reconstructed data. Black lines denote the training process of autoencoders. Colored lines denote different feature extraction paths when extracting feature groups.

- **Weakly**: This model tests our proposed weakly-supervised learning algorithm. Slice labels are not used during training.
- **Weakly-MultiRAED**: This model combines the proposed weakly-supervised learning program with the MultiRAED feature extraction. Slice labels are not used in training.
- **Supervised-MultiREAD**: Slices labels are used in training for this model. This serves as a upper bound for comparison with the weakly supervised learning.

Table 2. Accuracy on proprietary brain tumor dataset

Models	8-type	3-type
3D-DenseNet	$38.61\% \pm 7.88\%$	$55.00\% \pm 5.93\%$
3D-MultiRAED	$48.06\% \pm 13.67\%$	$63.33\% \pm 4.71\%$
Weakly	$47.23\% \pm 6.23\%$	$67.05\% \pm 7.09\%$
Weakly-MultiRAED	$56.33\% \pm 4.89\%$	$73.65\% \pm 3.65\%$
Supervised-MultiRAED	$\mathbf{57.13\% \pm 1.92\%}$	$\mathbf{73.95\% \pm 8.94\%}$

The classification results are shown in Table 2. We have the following observations: (1) Our two weakly supervised learning models are effective comparing to the 3D methods, which directly classify 3D volumes. The results of weakly supervised learning with MultiRAED is very close to the corresponding supervised version where slice-level labels are provided in training. (2) The effectiveness of MultiRAED is demonstrated by improvement of "3D-MultiRAED" compared to "3D-DenseNet", and "Weakly-MultiRAED" compared to "Weakly". MutliRAED improves feature learning with a large margin. 3) The similar result patterns can be observed in both 8-type and 3-type classifications. Other than the accuracy, the ROC curve is plotted in Fig. 3. Our dataset is much more challenging than the public dataset, where the accuracy has been pushed to more than 90%.

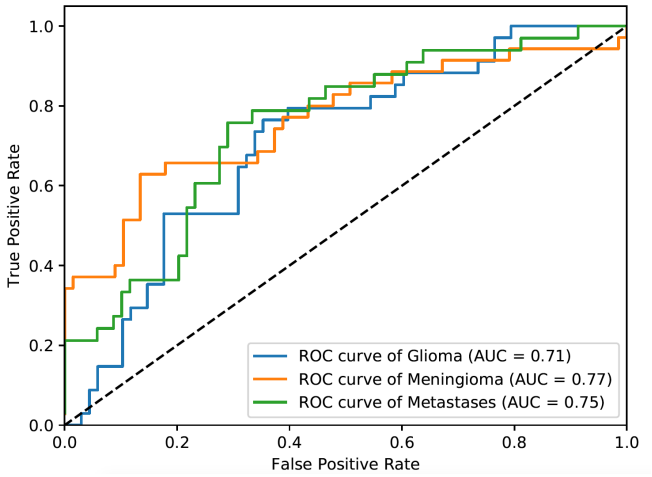

Fig. 3. ROC curve of 3-type tumor classification on our proprietary dataset, the figure is plotted based on one class vs. all other classes comparison.

4. CONCLUSION

We proposed a weakly supervised learning method without needing slice-level labels for efficient and effective tumor classification. The multiRAED is a general strategy to augment the limited training data in medical domain. The weakly supervised methods proposed here show a possible way of collecting large scale brain tumor dataset retrospectively without giving detailed slice level annotation.

5. REFERENCES

[1] Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes, "A survey of mri-based medical image analysis for brain tumor studies," *Physics in Medicine & Biology*, vol. 58, no. 13, pp. R97, 2013.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. Ginneken, and C. I.Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[3] Evangelia I Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R Melhem, and Christos Davatzikos, "Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme," *Magnetic resonance in medicine*, vol. 62, no. 6, pp. 1609–1618, 2009.

[4] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[5] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993, 2015.

[6] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," *PloS one*, vol. 10, no. 10, pp. e0140381, 2015.

[7] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis, "Brain tumor type classification via capsule networks," *arXiv preprint arXiv:1802.10200*, 2018.

[8] Yufan Zhou, Zheshuo Li, Hong Zhu, Changyou Chen, Mingchen Gao, Kai Xu, and Jinhui Xu, "Holistic brain tumor screening and classification based on densenet and recurrent neural network," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2018.

[9] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Qitao Huang, Muyan Cai, and Pheng-Ann Heng, "Weakly supervised learning for whole slide lung cancer image classification," *Medical Imaging with Deep Learning*, 2018.

[10] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache, "Deep learning with mixed supervision for brain tumor segmentation," *arXiv preprint arXiv:1812.04571*, 2018.

[11] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus, "Regularization of neural networks using dropconnect," in *International conference on machine learning*, 2013, pp. 1058–1066.

[14] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.