Orthogonal Ensemble Networks for Biomedical Image Segmentation

Agostina J. Larrazabal[1(✉)], César Martínez[1], Jose Dolz[2], and Enzo Ferrante[1]

[1] Research Institute for Signals, Systems and Computational Intelligence, Sinc(i), FICH-UNL/CONICET, Santa Fe, Argentina
alarrazabal@sinc.unl.edu.ar

[2] Laboratory for Imagery, Vision and Artificial Intelligence, École de Technologie Supérieure, Montreal, Canada

Abstract. Despite the astonishing performance of deep-learning based approaches for visual tasks such as semantic segmentation, they are known to produce miscalibrated predictions, which could be harmful for critical decision-making processes. Ensemble learning has shown to not only boost the performance of individual models but also reduce their miscalibration by averaging independent predictions. In this scenario, model diversity has become a key factor, which facilitates individual models converging to different functional solutions. In this work, we introduce Orthogonal Ensemble Networks (OEN), a novel framework to explicitly enforce model diversity by means of orthogonal constraints. The proposed method is based on the hypothesis that inducing orthogonality among the constituents of the ensemble will increase the overall model diversity. We resort to a new pairwise orthogonality constraint which can be used to regularize a sequential ensemble training process, resulting on improved predictive performance and better calibrated model outputs. We benchmark the proposed framework in two challenging brain lesion segmentation tasks –brain tumor and white matter hyper-intensity segmentation in MR images. The experimental results show that our approach produces more robust and well-calibrated ensemble models and can deal with challenging tasks in the context of biomedical image segmentation.

Keywords: Image segmentation · Ensemble networks · Orthogonal constraints

1 Introduction

In the past few years, deep learning-based methods have become the *de facto* solution for many computer vision and medical imaging tasks. Nevertheless, despite their success and great ability to learn highly discriminative features, they are shown to be poorly calibrated [1], often resulting in over-confident predictions. This results in a major problem, which can have catastrophic consequences in critical decision-making systems, such as medical diagnosis, where the downstream decision depends on predicted probabilities.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12903, pp. 594–603, 2021.
https://doi.org/10.1007/978-3-030-87199-4_56

Ensemble learning is a simple strategy to improve both the robustness and calibration performance of predictive models [2,3]. In this scenario, a common approach is to train the same model under different conditions, which can foster the model convergence to different functional solutions. Techniques to produce ensembles include dataset shift [4], Monte-Carlo Dropout [5], batch-ensemble [6] or different model hyperparameters [7], among others. Then, by averaging the predictions, individual mistakes can be dismissed leading to a reduced miscalibration. In this context, ensuring *diversity* across models is a key factor to build a robust ensemble. To promote model diversity in ensembles many mechanisms have been proposed. These include using latent variables [8], integrating attention in the embeddings to enforce different learners to attend to different parts of the object [9] or isolating the adversarial vulnerability in sub-models by distilling non-robust features to induce diverse outputs against a transfer attack [10].

Nevertheless, despite the relevance of obtaining well-calibrated models in clinical applications, relatively few works have studied this problem. Particularly, in the context of medical image segmentation, it was suggested that models trained with the well-known soft Dice loss [11] produce miscalibrated models [12], which tend to be highly overconfident. Furthermore, the recent work in [13] proposed the use of ensembles to improve confidence calibration. However, the importance of model diversity was not assessed in this work. Thus, given the negative impact of miscalibrated models in health-related tasks, and the current practices in medical image segmentation of systematically employing the Dice loss as an objective function, we believe it is of paramount importance to investigate the effect of ensemble learning in image segmentation, and how to enforce model diversity to generate high-performing and well-calibrated models.

Contributions. In this work, we propose a novel learning strategy to boost model diversity in deep convolutional neural networks (DCNN) ensembles, which improves both segmentation accuracy and model calibration in two challenging brain lesion segmentation scenarios. The main hypothesis is that inducing orthogonality among the constituents of the ensemble will increase the overall model diversity. We resort to a novel pairwise orthogonality constraint which can be used to regularize a sequential ensemble training process, resulting on improved predictive performance and better calibrated model outputs. In this context, our contributions are 3-fold: (1) we propose a novel filter orthogonality constraint for ensemble diversification, (2) we show that diversified ensembles improve not only segmentation accuracy but also confidence calibration and (3) we showcase the proposed framework in two challenging brain lesion segmentation tasks, including tumor and white-matter hyperintensity (WMH) segmentation on magnetic resonance images.

2 Related Works

Diversifying ensembles has been used to improve classification and segmentation performance of DCNNs in several contexts. In [14] authors propose an explicit way to construct diverse ensembles bringing together multiple CNN models and

architectures. Although they obtain successful results, this approach requires to manually design and train various architectures. An ensemble of 3D U-Nets with different hyper-parameters for brain tumor segmentation is proposed in [15], where authors point out that using different hyper-parameters reduces the correlations of random errors with respect to homogeneous configurations. However, no study on the diversity of the models and its influence on performance is presented. In [16] authors present a different view, highlighting that many automatic segmentation algorithms tend to exhibit asymmetric errors, typically producing more false positives than false negatives. By modifying the loss function, they train a diverse ensemble of models with very high recall, while sacrificing their precision, with a sufficiently high threshold to remove all false positives. While the authors achieve a significant increase in performance no study on the final calibration of the ensemble is carried out.

Following the success of ensemble methods at improving discriminative performance, its capability to improve confidence calibration has begun to be explored. [2] uses a combination of independent models to reduce confidence uncertainty by averaging predictions over multiple models. In [13] authors achieve an improvement in both segmentation quality and uncertainty estimation by training ensembles of CNNs with random initialization of parameters and random shuffling of training data. While these results are promising, we believe that confidence calibration can be further improved by directly enforcing diversity into the models instead of randomly initializing the weights.

As pointed out in [17] over-sized DNNs often result in a high level of overfitting and many redundant features. However, when filters are learned to be as orthogonal as possible, they become decorrelated and their filter responses are no longer redundant, thereby fully utilizing the model capacity. [18] follows a very similar approach but they regularize both negatively and positively correlated features according to their differentiation and based on their relative cosine distances. Differently from these works where ortogonality constraints are used to decorrelated the filters within a single model, here we propose to enforce filter orthogonality among the constituents of the ensemble to boost model diversity.

3 Orthogonal Ensemble Networks for Image Segmentation

Given a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})_i\}_{0 \leq i \leq |\mathcal{D}|}$ composed of images $\mathbf{x}$ and corresponding segmentation masks $\mathbf{y}$, we aim at training a model which approximates the underlying conditional distribution $p(\mathbf{y}|\mathbf{x})$, mapping input images $\mathbf{x}$ into segmentation maps $\mathbf{y}$. Thus, $p(y_j = k|\mathbf{x})$ will indicate the probability that a given pixel (or voxel) j is assigned class $k \in \mathcal{C}$ from a set of possible classes $\mathcal{C}$. The distribution is commonly approximated by a neural network $f_{\mathbf{w}}$ parameterized by weights $\mathbf{w}$. In other words, $f_{\mathbf{w}}(\mathbf{x}) = p(\mathbf{y}|\mathbf{x}; \mathbf{w})$. Parameters $\mathbf{w}$ are learnt so that they minimize a particular loss function over the training dataset. Given a set of segmentation networks $\{f_{\mathbf{w}^1}, f_{\mathbf{w}^2} ... f_{\mathbf{w}^N}\}$, a simple strategy to build an ensemble network $f_{\mathbf{E}}$ is to average their predictions as:

$$f_{\mathbf{E}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_{\mathbf{w}^i}(\mathbf{x}). \tag{1}$$

Under the hypothesis that diversifying the set of models $f^i_{\mathbf{w}}$ will lead to more accurate and calibrated ensemble predictions, we propose to boost its overall performance by incorporating pairwise orthogonality constraints during training.

Inducing Model Diversity via Orthogonal Constraints. Modern deep neural networks are parameterized by millons of learnable weights, resulting in redundant features that can be either a shifted version of each other or be very similar with almost no variation [18]. Inducing orthogonality between convolutional filters from the same layer of a given network has shown to be a good way to reduce filter redundancy [17]. Here we exploit this principle not only to avoid redundancy within a single neural model, but among the constituents of a neural ensemble.

Given two vectors $\mathbf{x}$ and $\mathbf{y}$, cosine similarity quantifies orthogonality (or decorrelation), ranging from -1 (i.e., exactly opposite) to 1 (i.e., exactly the same), with 0 indicating orthogonality. It can be defined as:

$$\mathrm{SIM}_C(\mathbf{x}, \mathbf{y}) = \frac{<\mathbf{x}, \mathbf{y}>}{||\mathbf{x}||\ ||\mathbf{y}||}. \tag{2}$$

Following [18], we consider the squared cosine similarity to induce orthogonality between filters through a new regularization term in the loss function. An advantage of this measure is that it takes into account both negative and positive correlations.

In order to enforce diversity within and between the ensemble models, we propose to include two regularization terms into the overall learning objective. The first one, referred to as self-orthogonality loss ($\mathcal{L}_{\mathrm{SelfOrth}}$), aims at penalizing the correlation between filters in the same layer, for a given model. Thus, for a given convolutional layer l, this term is calculated as follows:

$$\mathcal{L}_{\mathrm{SelfOrth}}(\mathbf{w}_l) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \mathrm{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}_{l,j})^2, \tag{3}$$

where $\mathbf{w}_{l,i}$ and $\mathbf{w}_{l,j}$ are vectorized versions of each of the n convolutional kernels from layer l. We also define an inter-orthogonality loss term ($\mathcal{L}_{\mathrm{InterOrth}}$) which penalizes correlation between filters from different models in the ensemble. To this end, following a sequential training scheme, the inter-orthogonality loss for layer l of model N_e is estimated as follows:

$$\mathcal{L}_{\mathrm{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}^e_l\}_{0\leq e<N_e}) = \frac{1}{N_e}\sum_{e=0}^{N_e-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}^e_{l,j})^2, \tag{4}$$

where $\{\mathbf{w}^e_l\}_{0\leq e<N_e}$ are the parameters of the previous $N_e - 1$ models trained during the sequential ensemble construction.

Thus, the learning objective to train the proposed OEN amounts to:

$$\mathcal{L} = \mathcal{L}_{Seg} + \lambda \sum_{l} \Big(\mathcal{L}_{\text{SelfOrth}}(\mathbf{w}_l) + \mathcal{L}_{\text{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}_l^e\}) \Big), \tag{5}$$

where $\mathcal{L}_{Seg}$ is the segmentation loss (e.g. soft Dice loss or cross entropy) and λ is a hyperparamether controlling the influence of the orthogonality terms.[1]

4 Experimental Framework

Database Description. We benchmark the proposed method in the context of brain tumor and WMH segmentation in MR images. For brain tumor we use the BraTS 2020 dataset [19–21] which contains 369 images with expert segmentation masks (including GD-enhancing tumor, peritumoral edema, and the necrotic and non-enhancing tumor core). Each patient was scanned with FLAIR, T1ce, T1, and T2. The images were re-sampled to an isotropic 1.0 mm voxel spacing, skull-stripped and co-registered by the challenge organizers. The provided training set, we divide the database in training (315), validation (17) and test (37). The second dataset [22] consists of 60 MR images with binary masks indicating the presence of WMH lesions. For each subject, co-registered 3D T1-weighted and a 2D multi-slice FLAIR images were provided. We split the dataset in training (42), validation (3) and test (15). All images have 3 mm spacing in the z dimension, and approximately 1 mm $\times$ 1 mm in the axial plane.

Segmentation Network. For all the experiments, the backbone segmentation network was a state-of-the-art ResUNet architecture [23] implemented in Keras 2.3 with TensorFlow as backend, with soft Dice [11] as segmentation loss $\mathcal{L}_{Seg}$. For the BraTS dataset, the input was a four-channel tensor (FLAIR, T1ce, T1, and T2) and a softmax activation was used as output, whereas a two-channel input (T1, FLAIR) was employed in the WMH, with a sigmoid activation function in the output. During training, patches of size $64 \times 64 \times 64$ were extracted from each volume, and networks were trained until convergence by sampling the patches randomly, with equal probability for each class in the case of tumour segmentation, and 0.9 probability in the case of WMH. We used Adam optimizer with a batch size of 64. The initial learning rate was set to 0.001 for BraTS and 0.0001 for WMH, and it was reduced by a factor of 0.85 every 10 epochs. Hyper-parameters were chosen using the validation split, and results reported on the hold-out test set.

Baselines and Ensemble Training. We trained two different baselines to benchmark the proposed method. In the first one (*random* ensemble) each model was randomly initialized and trained to reduce only the segmentation error $\mathcal{L}_{Seg}$. Therefore, its main source of diversity comes from the initialization of the weights. The second approach (*self-orthogonal* ensemble) includes the

[1] Our code associated to the orthogonal ensemble networks training is publicly available at: https://github.com/agosl/Orthogonal_Ensemble_Networks.

$\mathcal{L}_{\text{SelfOrth}}$ term in the learning objective, creating and ensemble of models individually trained with the self-orthogonality constraint. Thus, while each model learns orthogonal filters, orthogonality between different models in the ensemble was not imposed. We compared these two models with the proposed orthogonal ensemble network which also encourages inter-model diversity by minimizing the full objective defined in Eq. 5 (referred as *inter-orthogonal*). Note that in our approach models are trained sequentially. For each of the proposed settings we trained 10 models. During evaluation, we assembled groups of 1, 3 and 5 models from each setting by averaging the individual probability outputs. To provide better statistics, we repeated this process 10 times, each with different model selection. We empirically observed that beyond 5 models, the performance of the ensemble did not improve. Furthermore, λ was set to 0.1 and 1 for the WMH and brain tumour segmentation task, respectively.

Measuring Calibration for Image Segmentation. Given a segmentation network $f_{\mathbf{w}}$, if the model is well-calibrated its output for a single pixel j can be interpreted as the probability $p(y_j = k|\mathbf{x}; \mathbf{w})$ for a given class $k \in \mathcal{C}$. In this case, the class probability can be seen as the model confidence or probability of correctness, and can be used as a measure for predictive uncertainty at the pixel level [13]. A common metric used to measure calibration performance is the Brier score [24], a proper scoring rule whose optimal value corresponds to a perfect prediction. In other words, a system that is both perfectly calibrated and perfectly discriminative will have a Brier score of zero. In the context of image segmentation, for an image with N pixels (voxels), the Brier score can be defined as:

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{C}|}\sum_{k=1}^{|\mathcal{C}|}\Big(p(y_i = k|\mathbf{x}; \mathbf{w}) - \mathbb{1}[\bar{y}_i = k]\Big)^2, \tag{6}$$

where $\mathbb{1}[\bar{y}_i = k]$ is the indicator function whose value is 1 when $\bar{y}_i$ (the ground truth class for pixel i) is equal to k, and 0 otherwise.

Stratified Brier Score. In problems with highly imbalanced classes (such as brain lesion segmentation where most of the pixels are background), calibration may be good overall but poor for the minority class. In this case, the majority class will dominate and miscalibration in the class of interest will not be reflected in the standard Brier score. In [25], the authors proposed the stratified Brier score to measure calibration in binary classification problems with high imbalance. Here, we extend this concept to the segmentation task and propose to measure the stratified Brier score individually per-class, treating every structure of interest as a binary segmentation problem, to account for mis-calibration in the minority classes. For a given image with ground truth segmentation $\bar{\mathbf{y}}$, we construct the *stratified* Brier score for the class k, Br^k, by computing it only in the subset of pixels $\mathcal{P}_k = \{p : \bar{y}_p = k\}$, i.e. pixels whose ground truth label is k. The problem is therefore binarized considering all the other classes within a single background class. The formulation of the stratified Brier score Br^k is given by:

$$Br^k = \frac{1}{|\mathcal{P}_k|} \sum_{i \in \mathcal{P}_k} \Big(p(y_i = k|\mathbf{x}; \mathbf{w}) - \mathbb{1}[\bar{y}_i = k] \Big)^2 . \quad (7)$$

Segmentation Evaluation. In addition to the metrics presented to measure the model miscalibration, we resort to the common Dice Similarity coefficient (DSC) to assess the quality of the segmentations.

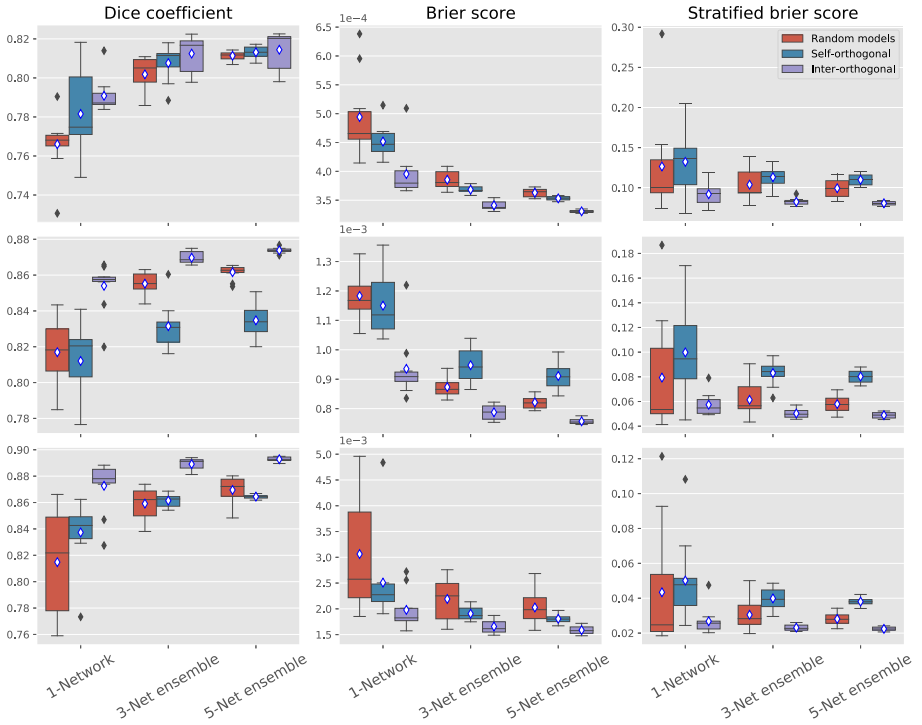

Fig. 1. Quantitative evaluation of the proposed method on BraTS: Rows from top to bottom show results for: (i) enhanced tumor; (ii) tumor core; (iii) whole tumor. Boxplots show mean and standard deviation for predictions obtained with individual models, 3-networks ensembles and 5-networks ensembles.

5 Results and Discussion

We present quantitative results for brain tumor and WMH segmentation in Fig. 1 and Fig. 2, respectively. We can observe that the model just integrating *self-orthogonality* outperforms the baseline model across groups and metrics. This improvement is further stressed when explicitly enforcing model diversity by

incorporating the $inter-orthogonality$ term computed between pairs of models during sequential training. In particular, our proposed learning strategy consistently leads to improvement on both model calibration and segmentation performance and across the two different segmentation tasks. This demonstrates the benefits of the proposed learning strategy to generate well-calibrated and highly performing segmentation models.

Another important observation is related to differences between Brier and stratified Brier scores. Given the small Brier value reported for all the models (less than 1^{-3}), one could think that these models are well calibrated. However, when having a closer look at the stratified Brier score, the higher value (more than 0.1 in most of the cases) reflects calibration issues. This results from the majority class dominating the traditional Brier score. Thus, studying the stratified Brier score allows us to better appreciate the improvements obtained by the inter-orthogonal ensemble with respect to the other models.

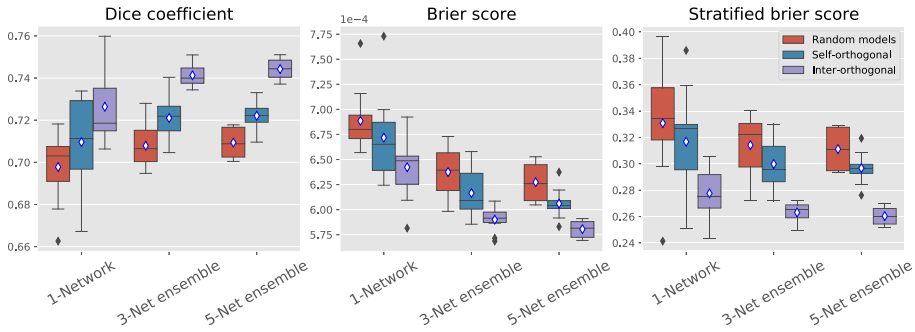

Fig. 2. Quantitative evaluation of the proposed method for WMH segmentation. Boxplots show mean and standard deviation for predictions obtained with individual models, 3-networks ensembles and 5-networks ensembles.

In addition, we depict in Fig. 3 the variance in the predictions across the components of the ensemble trained with and without the orthogonal losses, demonstrating that the orthogonal constraints bring diversity to the ensemble. As expected, we found that integrating the inter-orthogonal objective term leads to an increase in the variance of the predictions compared to the baseline models.

Last but not least, it is surprising to see that the inter-orthogonal regularization term boosts the performance even when considering the individual models. We believe that this is due to a regularization effect of the inter-orthogonal term, which implicitly reduces the complexity of the model by adding orthogonality constraints with respect to specific points in the parameter space, i.e. the weights of the previously trained models.

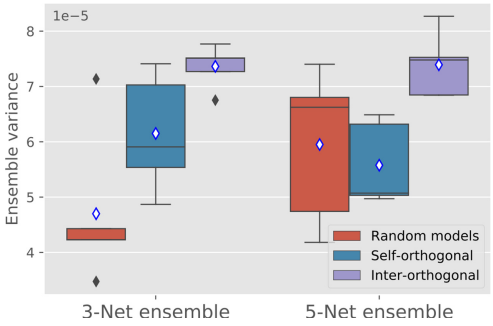

Fig. 3. Quantitative evaluation of the ensembles diversity. Boxplots depict the mean and standard deviation of the variance in the predictions when training the ensemble with and without the proposed orthogonal losses.

6 Conclusions

In this work we introduced Orthogonal Ensemble Networks (OEN), a novel training framework that produces more diverse ensembles. Our formulation explicitly imposes orthogonal constraints during training by integrating a regularization term that enhances the inter-model diversity. Experiments across two different segmentation tasks have demonstrated that, in addition to improved segmentation performance, the proposed inter-model orthogonality constraints reduce miscalibration, leading to more reliable predictions.

Acknowledgments. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-0620190100145LI, CAID-50220140100084LI) and ANPCyT (PICT 2018-03907). This research was enabled in part by support provided by Calcul Québec and Compute Canada.

References

1. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML, pp. 1321–1330 (2017)
2. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
3. Stickland, A.C., Murray, I.: Diverse ensembles improve calibration. In: ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning (2020)
4. Ovadia, Y., et al.: Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: NeurIPS (2019)
5. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
6. Wen, Y., Tran, D., Ba, J.: Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In: ICLR (2020)

7. Wenzel, F., Snoek, J., Tran, D., Jenatton, R.: Hyperparameter ensembles for robustness and uncertainty quantification. In: NeurIPS (2020)
8. Sinha, S., Bharadhwaj, H., Goyal, A., Larochelle, H., Garg, A., Shkurti, F.: Dibs: diversity inducing information bottleneck in model ensembles. In: AAAI (2020)
9. Kim, W., Goyal, B., Chawla, K., Lee, J., Kwon, K.: Attention-based ensemble for deep metric learning. In: ECCV, pp. 736–751 (2018)
10. Yang, H., et al.: DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles. arXiv preprint arXiv:2009.14720 (2020)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 IEEE (2016)
12. Sander, J., de Vos, B.D., Wolterink, J.M., Išgum, I.: Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In: Medical Imaging 2019: Image Processing, vol. 10949, p. 1094919. International Society for Optics and Photonics (2019)
13. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging **39**(12), 3868–3878 (2020)
14. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 450–462. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_38
15. Feng, X., Tustison, N.J., Patel, S.H., Meyer, C.H.: Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features. Front. Comput. Neurosci. **14**, 25 (2020)
16. Ma, T., et al.: Ensembling low precision models for binary biomedical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 325–334 (2021)
17. Wang, J., Chen, Y., Chakraborty, R., Yu, S.X.: Orthogonal convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11505–11515 (2020)
18. Ayinde, B.O., Inanc, T., Zurada, J.M.: Regularizing deep neural networks by enhancing diversity in feature extraction. IEEE Trans. Neural Networks Learning Syst. **30**(9), 2650–2661 (2019)
19. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
20. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
21. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
22. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans. Med. Imaging **38**(11), 2556–2568 (2019)
23. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-Net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)
24. Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **78**(1), 1–3 (1950)
25. Wallace, B.C., Dahabreh, I.J.: Improving class probability estimates for imbalanced data. Knowl. Inf. Syst. **41**(1), 33–52 (2013). https://doi.org/10.1007/s10115-013-0670-6