

Distributing Deep Learning Hyperparameter Tuning for 3D Medical Image Segmentation

Josep Ll. Berral[†], Oriol Aranda, Juan Luis Dominguez, Jordi Torres[†]
Barcelona Supercomputing Center, Universitat Politècnica de Catalunya[†]
josep.berral@bsc.es, oriol.aranda@bsc.es, juan.dominguez@bsc.es, jordi.torres@bsc.es

Abstract—Most research on novel techniques for 3D Medical Image Segmentation (MIS) is currently done using Deep Learning with GPU accelerators. The principal challenge of such technique is that a single input can easily cope computing resources, and require prohibitive amounts of time to be processed. Distribution of deep learning and scalability over computing devices is an actual need for progressing on such research field. Conventional distribution of neural networks consist in "data parallelism", where data is scattered over resources (e.g., GPUs) to parallelize the training of the model. However, "experiment parallelism" is also an option, where different training processes (i.e., on a hyper-parameter search) are parallelized across resources. While the first option is much more common on 3D image segmentation, the second provides a pipeline design with less dependence among parallelized processes, allowing overhead reduction and more potential scalability. In this work we present a design for distributed deep learning training pipelines, focusing on multi-node and multi-GPU environments, where the two different distribution approaches are deployed and benchmarked. We take as proof of concept the 3D U-Net architecture, using the MSD Brain Tumor Segmentation dataset, a state-of-art problem in medical image segmentation with high computing and space requirements. Using the BSC MareNostrum supercomputer as benchmarking environment, we use TensorFlow and Ray as neural network training and experiment distribution platforms. We evaluate the experiment speed-up when parallelizing, showing the potential for scaling out on GPUs and nodes. Also comparing the different parallelism techniques, showing how experiment distribution leverages better such resources through scaling, e.g. by a speed-up factor from x12 to x14 using 32 GPUs. Finally, we provide the implementation of the design open to the community, and the non-trivial steps and methodology for adapting and deploying a MIS case as the here presented.

Index Terms—Distributed Deep Learning, Distributed Computing, GPU, Parallelism, Scalability

I. INTRODUCTION

In the past decade, Deep Learning methods (DL) have revolutioned the fields of machine learning, computer vision, and data analytics. This has supposed a huge leap forward in academic research, also in industrial development. Medicine data analysis is one of these fields, leveraging state-of-art Neural Networks for 3D Medical Image Segmentation (MIS), to create models with ground-breaking accuracy and efficiency. There is a large catalog of DL techniques and models focusing on detection, diagnosis and segmentation of 3D data. The principal challenge is that processing 3D medical data on neural networks is computationally expensive and requires high amounts of memory space. Speeding-up approaches attempt to split data in subpatches before feeding them to

the deep network, reducing memory usage requirements but losing spatial information required for good accuracy. In cases targeting the highest accuracy, 3D images need to be processed as full images, then distribution and parallelism are a necessity.

Adapting and tuning state-of-art DL models for Medical Image Segmentation is not trivial. A single model training experiment can be costful in resources and time, not to mention that in most cases hyper-parameter search is also required, multiplying the number of training processes to perform. Scaling out computing resources becomes a necessity, by parallelizing the different experiments and modeling processes across the available resources as computing nodes and GPUs. Distribution of the different parts of the training pipeline, including data transformation, data deployment and process placement, must be properly engineered and adapted. Distributed deep learning pipelines can be designed towards "data distribution", where data is distributed along different GPUs for training models as a All-Reduce process, or "experiment distribution", where different GPUs deal with different models in parallel. Previous works explored how data distribution can speed up DL training; however, both approaches can be used to speed-up sets of DL training experiments.

In this work we present the design, methodology and evaluation of Distributed Deep Learning approaches, considering multi-node and multi-GPU environments for scaling out resources. As proof of concept and benchmarking use case, we distribute a state-of-art full-3D volume Medical Image Segmentation network, the 3D U-Net for Brain Tumor Segmentation, on the BSC MareNostrum supercomputing GPU environment. Our methodology pays special attention on data transformation from standard MIS formats (i.e., NIfTI [1], DICOM [2]), and uses DL and experiment distribution frameworks like TensorFlow and Ray for both data and experiment parallelism. We evaluate the scalability of the different distribution methods, proving the potential for multi-node and multi-GPU scale out, and comparing the speed-up provided by both distribution methods.

Acceleration of DL pipelines on Medical Image Segmentation are an important medical research use case, in which supercomputing centers are getting more involved day by day. Supercomputing research centers have available large pools of resources to research on High-Performance Computing, including novel architectures and platforms, optimization of high-performance demanding applications, and accelerating Artificial Intelligence and DL use cases. A case like the pre-

sented Brain Tumor Segmentation, using 3D U-Net networks with full volume input, is a modelic benchmark for such research where the pipeline distribution is not trivial. Full volume input requires heavy memory usage offered by GPUs, and data must be transformed and arranged for fitting in the device without disrupting the pipeline. And it is known that alternative shortcuts for treating this problem, like subpatching the input dataset, do not perform as good as desired due to the loss of spatial information. Furthermore, full-volume input converges faster, reducing training and inference time.

The evaluation of the presented approaches, and the study of DL distribution scalability are performed in the MareNostrum-CTE GPU Supercomputing environment, composed by state-of-art GPUs NVIDIA V100 16GB deployed on a grid of HPC computing nodes. Results on model dice score (DSC) are kept as reference, to ensure that any pipeline or data modification affects the quality of the resulting models. As a result, we provide the times and speed-up for resource scaling out between 1 GPU and 32 GPUs, in 4-GPU computing nodes. We also observe that in experiment parallelism, distributed components have less dependence than data parallelism, introducing less overhead when scaling out, increasing speed-up from $\times 12$ to $\times 14$ with respect data parallelism on 32 GPUs. Finally, the implementation of the presented methodology, with all the details and non-trivial adjustments required for adapting the standard 3D U-Net model and the MSD Brain Tumor dataset, is made public and available to the community, as means to help deploying such workloads on scientific clusters, supercomputers and Cloud environments.

Summarizing, the contributions presented in this paper are:

- 1) The design and methodology for distributing Deep Learning MIS towards full-3D volume input, considering data versus experiment distribution with hyperparameter search, across multi-node multi-GPU environments.
- 2) The adaption of the Brain Tumor Segmentation using 3D U-Net network, as benchmark of speed-up on multi-node and multi-GPU environments. Also, the study and comparison of the speed-up offered by the different distribution approaches.
- 3) A framework using TensorFlow and Ray for training and inference on multi-GPU environments, considering the non-trivial adaptations for full volume 3D imaging, as guide and community open-source software.

This paper is organized as follows: Section II introduces the state of the art and background methodologies. Section III describes the presented methodology. Section IV shows the experiments and results validating our approach. Finally, Section V provides some discussion and the conclusions.

II. STATE OF THE ART

A. Related Work

1) *Deep Learning on Medical Image Segmentation*: Since introduction of the U-Net in [3], this type of network and its variants have achieved state-of-the-art results on various

2D and 3D medical image segmentation tasks [4], [5], [6], [7]. Especially in 3D image brain tumor segmentation the 3D U-Net model [8], shown to be better for volumetric medical data, is widely used [9], [10]. Numerous approaches of this model have been developed using sampled sub-volume patches [11], [12], because of memory limitations and to focus more precisely on tumor regions. This last approach, despite it leads to good qualitative and quantitative results, loses spatial information and it has very poor performing time for both training and inference. Our proposed approach is an end to end solution using the full volume input, thus spatial information is not lost, leads to good qualitative results but also better convergence time and hence better scalability.

2) *Distributed Deep Learning*: In previous works [13], we explored how distributed learning can help to speed up training for neural networks. Several work on spatial, model and data parallelism has been done during the recent years [14], [15], [16], including the implementation of these techniques into the most used deep learning frameworks, e.g. Tensorflow. In spatial parallelism, the data is split into subpatches, and they are sent into different devices with some spatial information. Model parallelism consist on splitting the model and each device is responsible for computing its own piece. In data parallelism, a batch of data is split across the devices and each one computes a mini-batch; it's the most used and is demonstrated as the most efficient and preferred approach whereas either the model or a sample of data can be fed into memory. All these approaches try to solve the common problem of memory limitations when using heavy datasets or models, hence, specially in medical images their application has been also studied [17]. Spatial parallelism has been applied for high resolution medical image analysis [18]. Model parallelism has been proposed for medical image segmentation [19]. Data parallelism has been also used for COVID-19 diagnosis based on CT scans [20], and text and feature extraction diagnosis using CNN models [21], [22]. Some studies combine some of the abovementioned techniques, called hybrid parallelism, to handle 3D images and models [23], [24]. In [25] a scalable toolkit for medical image segmentation is presented, but is privative and only two models are provided. In our method we use a data parallelism approach, and we integrate the whole pipeline for preprocessing and reading the data which is crucial for a good training performance.

3) *Distributed Hyperparameter tuning*: Since the beginning when the use of neural networks was first introduced, the hyperparameter optimization or tuning has been essential to improve their performance [26]. It is also well known, that it is a tedious and slow process, for that reason several studies on distributing it and thus increase its performance have been carried out [27], [28]. Furthermore, it has also been proposed on medical image diagnosis [29], [30], [31], but in these studies their focus is not on efficiency. In our work we propose an easy to use distributed hyperparameter tuning, which leads to a dramatically improvement on performance and more simple usability.

B. Background

1) *The 3D U-Net Model*: The 3D U-Net [8] is the most used model for segmentation tasks in medical imaging. At a high level, the network has an analysis and a synthesis path, the encoder and the decoder respectively with four resolution steps each. In the analysis path, each layer contains two $3 \times 3 \times 3$ convolutions each followed by a rectified linear unit (ReLU), and then a $2 \times 2 \times 2$ max pooling with strides of two in each dimension. In the synthesis path, each layer consists of a transposed convolution of $2 \times 2 \times 2$ by strides of two in each dimension, then a concatenation layer followed by two $3 \times 3 \times 3$ convolutions each followed by a ReLU. Shortcut connections (concatenations) from layers of equal resolution in the analysis path provide the essential high-resolution features, i.e. spatial information from early layers, to the synthesis path. Referring to the analysis path, the number of filters used in both convolutions at each resolution step are doubled. In turn, the number of filters for the synthesis path is halved.

2) *Loss Function for Segmentation*: Aside from the architecture, one of the most important elements of any deep learning method is the choice of the loss function. Due to heavy class imbalance (there are typically not many positive regions) the Dice similarity coefficient¹ (DSC), which is a measure of how well two contours overlap, is commonly used. Given A and B as sets of voxels, A being the predicted tumor region and B being the ground truth, the Dice index ranges from 0 (complete mismatch) to 1 (perfect match). The model outputs probabilities that each pixel is a tumor or not, and those outputs are desired to be backpropagated through. Therefore, an analogue of the Dice index which takes real valued input is utilised [32], [33]. Additionally, the loss function is minimized during training, so it is defined to decrease as performance increases:

$$\mathcal{L}_{Dice}(\hat{y}, y) = 1 - \frac{2 \times \sum_{i,j} \hat{y}_{ij} y_{ij} + \epsilon}{\sum_{i,j} \hat{y}_{ij} + \sum_{i,j} y_{ij} + \epsilon}$$

Where \hat{y} is the prediction mask, y the ground truth mask and ϵ is a small constant, i.e. 0.1, added to avoid division by zero. Another variant of the loss, called quadratic Soft Dice Loss following [7], is tested along with the dice coefficient as the metric but seems to lead to worst validation results.

3) *TensorFlow and Ray API*: TensorFlow and Ray provide APIs to define the training and validation pipelines, towards optimizing the data encoding and distribution, also the experiment definition for its distribution. Experiment distribution will split the hyper-parameter search across the working nodes, performing all the same modelling process with different tuning. On the other hand. Data distribution will split the dataset among the working nodes, each node modelling its sub-set and synchronizing updates among nodes.

In Tensorflow, *tf.Data* [34] provides optimization for pipelines, transparent to the users, allowing also to preprocess data in parallel. E.g., the transformation of raw data into

TFRecords, the optimized internal format for TensorFlow data, is directly parallelized through the “interleave” and “map-reduce” functions. Also, *tf.MirroredStrategy* [35] provides data parallelism across devices on the same machine, creating replicas or copies of a model to be run on different slices of the input data. Note that when using data parallelism, the batch size is divided across devices. Then, to take full advantage of the available devices, the batch size (and the initial learning rate) must be multiplied by the number of devices used.

For Ray, data parallelism is achieved through *Ray.Cluster* [36] and *Ray.SGD* [37] libraries, but over multiple machines instead of devices. Ray handles all the communication between nodes. In addition, *Ray.Tune* is in charge of performing distributed hyperparameter tuning over the most popular machine learning frameworks. With this, the researcher only needs to focus on the training settings, letting Ray to handle experiment distribution.

III. METHODOLOGY

The proposed methodology focuses on two main scenarios: parallelizing data for training, and parallelizing hyperparameter tuning, both distributing the workload in the two different ways aforementioned. Figure 1 presents the schema of the created pipeline for the two different approaches.

First approach is the distribution of the training process across multiple computing devices (here GPUs, across a multi-node HPC cluster), leveraging the *Distributed TensorFlow* API to split the experiment data batches across the available resources in each node, then *Ray.SGD* for distributing across nodes. The second approach focuses on the parallelization of the hyper-parameter tuning, using *Ray.Tune* [38] to efficiently distribute the different experiments on hyper-parameter combinations. Both approaches use the training and validation datasets for training and evaluation respectively on each model, also they explore the same combination of hyper-parameters.

A. Model Specification

The model used for benchmarking the presented set-up is the previously introduced 3D U-Net model. As shown in Figure 2, the number of filters for each resolution step $s \in \{1, 2, 3, 4\}$ is $8 \times 2^{s-1}$. Further, a $1 \times 1 \times 1$ convolution followed by a sigmoid reduces the number of output channels in the last layer, to match the number of output labels in our case (i.e., 1). Finally, we used *batch normalization* before each ReLU, and a truncated normal kernel initializer for each convolution layer.

As data format, the neural network is built with *Channels First*, being the input a $4 \times 240 \times 240 \times 152$ voxel tile of the image, and the output a $1 \times 240 \times 240 \times 152$ voxel tile matching the ground truth. Once compiled, our benchmarking neural network has 406.793 parameters in total.

B. Architecture Details

1) *Adapting the Pipeline*: Either using Ray.Tune for distributing hyper-parameter search or not using it, the used Neural Network engine is TensorFlow (TF). TF uses binarized

¹Dice similarity coefficient is known by several names such as Sørensen-Dice Coefficient, Dice’s coefficient, Dice index or F1-score

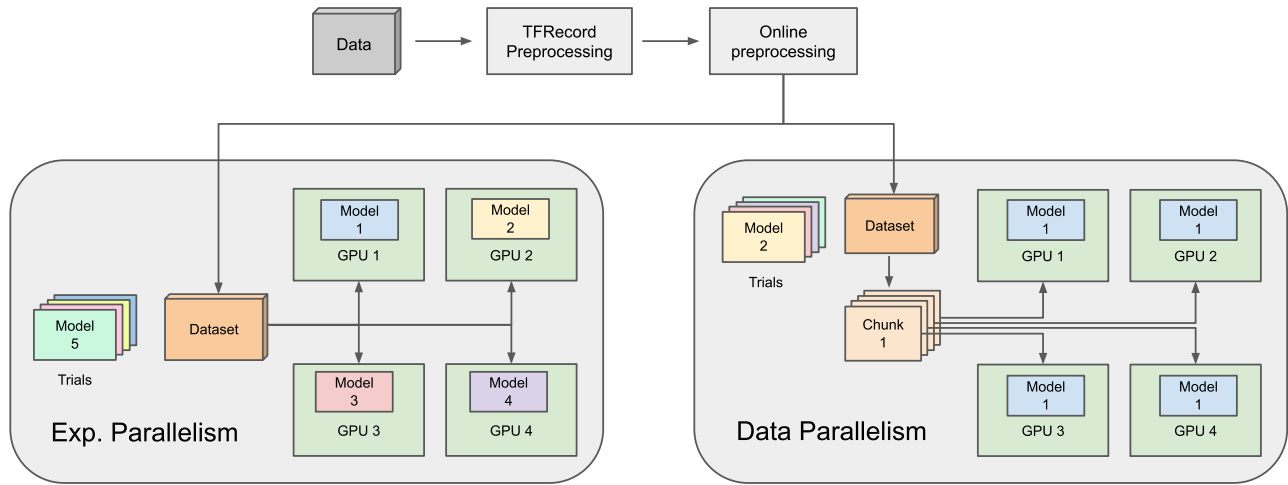


Fig. 1: Schema of the different approaches on the proposed methodology

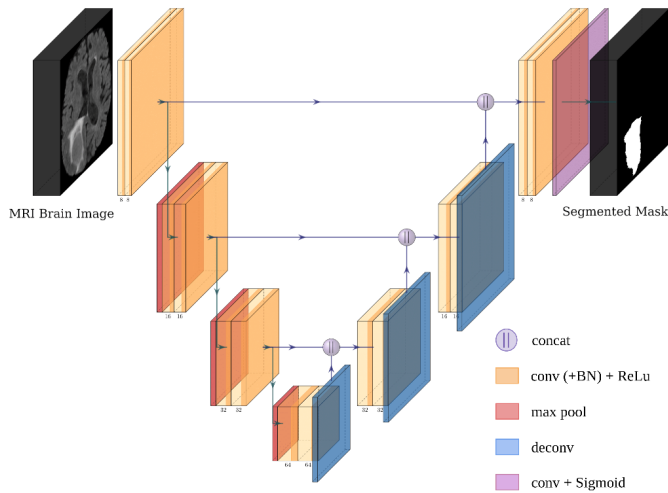


Fig. 2: Overview of the used 3D U-Net architecture.

records as input, i.e. the *TFRecord* format [39], converting images to binary data during the training phase. An initial analysis on test runs, using the Tensorboard profiler tool [40], [41], showed us that data loading and its transformation into binary records are the principal bottlenecks in the preprocessing stage of the pipeline, something totally expected from the size and complexity of the target data. Knowing that the input data will remain the same after each epoch, such data can be binarized off-line before starting the training process. This way, we can avoid to pre-process the data at each epoch, reducing significantly the training cost. Reading the files for binarization can be parallelized using interleaved functions, while the binarization process can be mapped over the read data. In addition, the dataset can be pre-fetched.

2) *Parallelism Levels*: On both approaches, data and experiment parallelism, a hyper-parameter space must be defined with any desired configuration to be performed. In our scenario, this set of configurations becomes the cross-product of the different values for each option in the configuration.

Then, for the data parallelism approach, we can identify three cases when training, depending on the number of available GPUs in the system as n :

- $n = 1$: Since we only have 1 GPU, training becomes sequential, and each experiment is produced sequentially without any parallelism.
- $1 < n \leq M$: Consider M as the number of GPUs on a single computing node ($M = 4$ in our scenario). Here, the Distributed Tensorflow API is triggered to parallelize up to M experiments in parallel in the single computing node.
- $n > M$: As more than a single computing node is used, Ray.Cluster is launched to "create" a cluster of available and reachable resources across physical nodes. With Ray.SGD we apply data parallelism across multiple machines, also Ray handles all the intercommunications and model training updates between nodes.

Finally, for the experiment parallelism approach, Ray.Cluster is directly launched to "create" a parallelism cluster along the available resources, then Ray.Tune performs the distributed hyper-parameter tuning process. Adapting any neural network to Ray.Tune implies adapting its implementation to the standard Ray API, or set of functions for fit, evaluate and predict that Ray expects to find and execute. The basic requirements are to have the training process in a "training" function to be called from Ray, having a dictionary containing the hyperparameters as argument. Also, a reporting callback function is required, to provide Ray

with the finalization results. Then, the batch of experiments are run through Tune.Run, passing the set of hyper-parameters to explore.

IV. EXPERIMENTS

A. Dataset

The dataset used as benchmarking for these sets of experiments is the "Task 1" dataset (brain tumor MRI segmentation), from the MSD challenge [42]. Such dataset consists on 484 multi-modal multi-site MRI data (FLAIR, T1w, T1gd T2w) with 4-class ground truth labels referring to "background", "enhancing" and "non-enhancing tumor", and "edema" segmentations. The volume size for each image is [240, 240, 155], and its resolution/spacing is uniformly $1.0 \times 1.0 \times 1.0 \text{ mm}^3$. Also, for MRI images, the voxel intensities are pre-processed through standardization. Figure 3 shows a sample of the original dataset before pre-processing towards benchmarking, i.e. the four channels and the ground truth image.

The problem corresponding to the original dataset corresponds to a 4-class classification, but for our benchmarking we are reducing the problem to a binary class segmentation task (whole tumor vs. background). The three non-background classes are joined in a single label for "positive", while the background label is considered as "negative". Because of the model architecture, the input dataset must be cropped to sizes of [240, 204, 152] and transposed to "channel first" (from 4-channel inputs) data format. Finally, the dataset is split for training, validation and evaluation as 70%, 15% and 15% respectively.

B. Deployment and Implementation

The 3D U-Net model is implemented in TensorFlow 2.3, and Ray 1.4.1 is used for the hyper-parameter tuning. The benchmarking pipeline has been deployed in the Barcelona Supercomputing Center MareNostrum-CTE cluster, composed by 52 IBM Power9 (8335-GTH @2.4GHz CPUx20) nodes with 4 NVIDIA V100 16GB GPUs each. Infiniband is used as interconnection network. Scalability has been tested for 1 to 32 GPUs (1 to 8 machines). These are the modules which build the stack of software: gcc/8.3.0 cuda/10.2 cudnn/7.6.4 nccl/2.4.8 python/3.7.4.

One of the principal challenges on these deployments is the volume of each input, against the capacity of the available computing devices. State-of-art GPUs, although having enough resources for common problems, still have very limited memory for use cases like the 3D U-Net. The model is trained with a batch size of 2 inputs per replica, meaning a total batch of $2 \times \# \text{GPUs}$ when an experiment is distributed. The number of epochs per experiment is 250, although the model converges much faster and both training and validation are stabilized around epoch 90. Even when distributing data across devices, convergence does not differ much when scaling up, as the training process is synchronized through SGD. The optimizer algorithm used is Adam [43], with an initial learning rate of $10^{-4} \times \# \text{GPUs}$. Notice here that the learning rate depend on the ratio of data distribution, and because of the scattering of

data batches across devices, we need to approximate it, e.g., by using the Cyclic Learning Rates technique [44]. Aside of the Learning rate (values: $1e-04$, $5e-04$, $1e-03$, $5e-03$), the explored parameters include the normalization method (batch vs. group normalization), the use of AMSGrad, and the ADAM b_2 (0.99, 0.999). This choice of hyperparameter grid provides 32 different combinations, allowing us to test the experiment parallelism from 1 to 32 GPUs.

C. Performance Analysis

In order to validate our study and comparison benchmarking, we must ensure that the proposed architecture, deployment and modifications do not affect the performance of the models in terms of correctness (i.e., dice score). For the different experiments here performed, the evaluation on the validation and test sets provide a dice score of 0.89, which are the results of the state-of-art 3D U-Net model. Hence, our methodology and architectures are capable of keeping the dice score results while improving the performance notably. As an example of the evolution of the dice score along epochs, Figure 4 shows convergence starting at 20 epochs, per number of GPUs on data parallelism also on experiment parallelism.

The following comparison between the two distribution architectures, as seen in Table I, shows the scalability and speed-up provided by doubling the resources (i.e. GPUs), from 1 to 32. Every execution has been run three times, providing here the average.

# GPUs used	Data Parallel Method		Experiment Parallel Method	
	Elapsed time	Speedup	Elapsed time	Speedup
1	44:18:02	1.00	44:20:19	1.00
2	23:09:28	1.91	22:24:39	1.98
4	15:09:35	2.92	11:32:20	3.84
8	7:41:12	5.76	7:03:17	6.28
12	5:59:59	7.38	5:35:22	7.93
16	4:26:50	9.96	4:11:54	10.56
32	3:21:44	13.18	2:55:06	15.19

TABLE I: Results on data parallelism method and experiment parallelism method

We observe that both architectures follow an almost-linear progression on speed-up when doubling the available GPUs. We must have into account that every computing node has 4 GPUs, and using more implies a communication overhead when distributing a single model across nodes. In comparison, when distributing the hyper-parameter search, each execution is independent from the next one. That prevents overheads from data shuffling or intermediate results communication, as every parallel run is self-contained.

Figure 5 displays those results, highlighting the difference between both methods: on the average elapsed time per number of GPUs, and the average speed-up per number of GPUs. Although both methods scale really well, the Ray.Tune method for hyper-parameter distribution shows better improvement on time/speed-up. It is important to recall that, given the amount of experiments that are usually performed when finding the best model for MRI, the smallest improvement on execution time for these kind of experiments can easily add up to

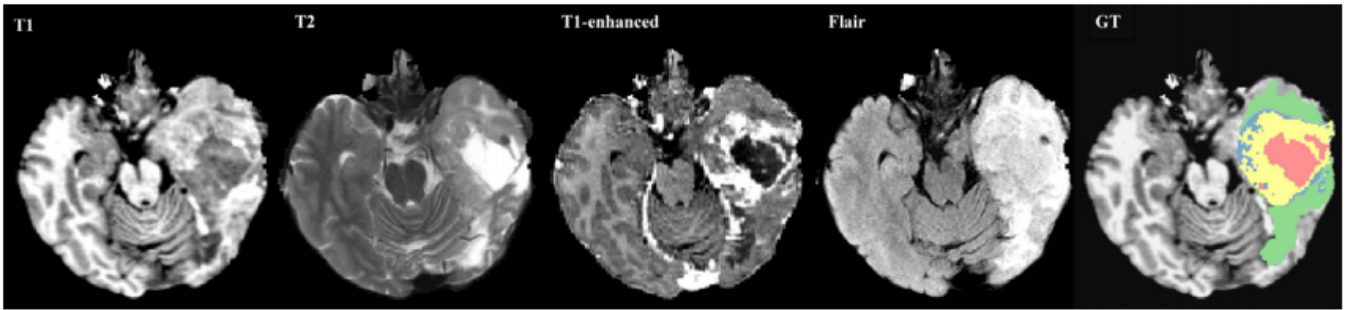


Fig. 3: Overview of one sample of the data. The first 4 images corresponds, from left to right, T1w (T1-weighted), T2w (T2-weighted), T1gd (T1-weighted with gadolinium contrast enhancement) and FLAIR (Fluid Attenuated Inversion Recovery). The last corresponds to the ground truth.

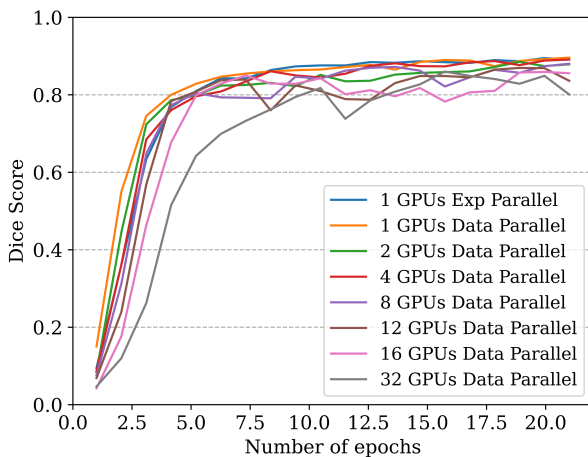


Fig. 4: Dice Score evaluation per number of GPUs in data parallelism and experiment parallelism

hours and days when repeating runs or expanding the hyper-parameter search space.

V. CONCLUSIONS

A. Summary

In this work we are proposing, studying and evaluating the distribution and scalability of heavy-data workloads, like Neural Network training for Medical Image Segmentation, on multi-GPU and multi-node architectures. This kind of data and problems pose a problem when being deployed on fast-computing devices like GPUs, due to their resource demant and memory occupation of each single input image. Cases like MIS for brain tumor using state-of-art neural networks like 3D U-Net make evident the urgent need for scaling for medical research to leverage deep learning.

Deploying a multi-node multi-GPU architecture provides different options for distributing neural network experiments: data distribution, where data is scattered across available devices to process them in parallel for a single model; and experiment distribution, where each device trains different models

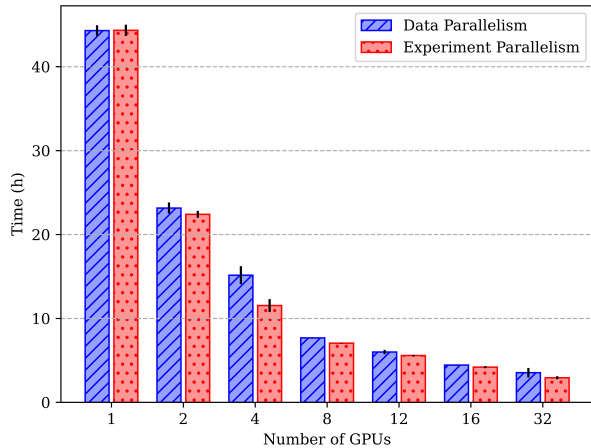
from a list of potential model configurations (i.e., hyper-parameter search). Here we provide the system design for both kinds of distribution, with the practical details when deploying a MIS 3D U-Net network as a proof of concept. We highlight the procedures and steps to adapt the problem towards a multi-node multi-GPU cluster (e.g., the BSC MareNostrum supercomputer), and we provide the implementation as open source software as reference for researchers and engineers that require deploying such pipelines.

After benchmarking the proposed architectures using the MSD Brain Tumor Segmentation dataset, with the 3D U-Net image segmentation network, we show the potential of scaling multi-device and node. Also, we compare the two different experiment distribution approaches, in a hyper-parameter search scenario, by serializing experiments while doing data distribution, and by distributing single-device experiments as experiment distribution. This has been done using TensorFlow and Ray platforms for Neural Network training and experiment parallelism.

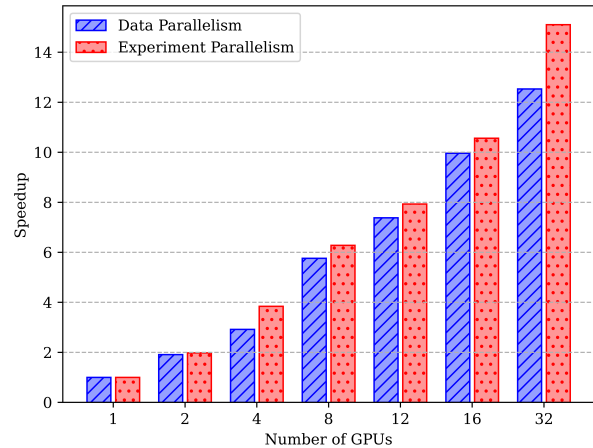
B. Open Community Framework

As initially mentioned, one of the main contributions is to provide the results of this work as an open framework and guide for the community². The principal efforts of this work have been set in finding the correct configurations and capabilities of large-scale computing systems, to show direct improvement of such Deep Learning distribution approach, in a way that these and other use cases can leverage. The idea behind this is that those users who require Deep Learning in any medical imaging task, can proceed efficiently by following the deployment guide on their systems, and adapt the framework for their purposes (e.g., changing the model architecture and the dataset reader to their desired ones, etc). This will allow users to use the properly integrated pipeline, with the option to automatically perform DL distribution as hyper-parameter tuning or as single experiment training. In order to allow a broader range of users to use this work, there is also available a PyTorch version of the code in the repository.

²Available Open-Source code: <https://github.com/HiEST/DistMIS> (Mar'22)



(a) Average elapsed time per number of GPUs, with max and min



(b) Average speed-up per number of GPUs

Fig. 5: Comparison of mean elapsed time (a) and speedup (b) between the two methods.

C. Future Work

An important issue, noticed during the scaling over GPU acceleration devices, is the limitation of memory when processing datasets with large inputs, as happens in 3D MIS learning. On scenarios like that, batch sizes are forcefully reduced to 2 or even 1 input, as there is no room in GPU memory for more. A scale-out solution to such problems is to consider model or pipeline parallelism, where the training pipeline for a single model is split across devices, that is, additional GPU devices are used to fit the memory requirements. Such distribution is more difficult than data or experiment distribution, since the neural network must be disaggregated. However, models like 3D U-Net, that involve skipping connections, cannot be trivially used by these schemes and applying such approach would require a deeper inspection. Other techniques that could allow to scale up model size in single GPU devices include CPU offloading of the model states, mixed precision training, and activation checkpointing. Next steps focus on scaling resources using model and pipeline parallelism, and scaling up techniques, to surpass the problem of large input units. Frameworks allowing to distribute models are becoming state-of-art, and being pushed on by GPU manufacturers (e.g., NVidia and DeepSpeed [45]) showing the potential of such techniques and devices to accelerate Deep Learning even more.

ACKNOWLEDGEMENTS

This work has been partially financed by the European Commission (EU-H2020 INCISIVE GA.952179, and CALLISTO GA.101004152). Also the Spanish Ministry of Science (PID2019-107255GB-C22/ AEI / 10.13039/501100011033), and Generalitat de Catalunya through the 2017-SGR-1414 project.

REFERENCES

[1] National Institute of Mental Health, “Nifti documentation,” October 2021. Available at <http://nifti.nimh.nih.gov/nifti-1/documentation>.

[2] P. Mildenerger, M. Eichelberg, and E. Martin, “Introduction to the dicom standard,” *European radiology*, vol. 12, no. 4, pp. 920–927, 2002.

[3] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer, 2015.

[4] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, “U-net and its variants for medical image segmentation: A review of theory and applications,” *IEEE Access*, 2021.

[5] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “Unet 3+: A full-scale connected unet for medical image segmentation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[6] S.-T. Tran, C.-H. Cheng, T.-T. Nguyen, M.-H. Le, and D.-G. Liu, “Tmd-unet: Triple-unet with multi-scale input features and dense skip connection for medical image segmentation,” in *Healthcare*, vol. 9, p. 54, Multidisciplinary Digital Publishing Institute, 2021.

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. 2016.

[8] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: Learning dense volumetric segmentation from sparse annotation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), (Cham), pp. 424–432, Springer, 2016.

[9] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, “A variant form of 3d-unet for infant brain segmentation,” *Future Generation Computer Systems*, vol. 108, pp. 613–623, 2020.

[10] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, “A two-stage 3d unet framework for multi-class segmentation on full resolution image,” *arXiv preprint arXiv:1804.04341*, 2018.

[11] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, “Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge,” in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, eds.), (Cham), pp. 287–297, Springer, 2018.

[12] J. Chang, X. Zhang, M. Ye, D. Huang, P. Wang, and C. Yao, “Brain tumor segmentation based on 3d unet with multi-class focal loss,” in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, IEEE, 2018.

[13] V. Campos, F. Sastre, M. Yagües, M. Bellver, X. G. i Nieto, and J. Torres, “Distributed training strategies for a computer vision deep learning algorithm on a distributed gpu cluster,” in *ICCS*, 2017.

- [14] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.
- [15] Z. Jia, M. Zaharia, and A. Aiken, "Beyond data and model parallelism for deep neural networks," 2018.
- [16] C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl, "Measuring the effects of data parallelism on neural network training," *Journal of Machine Learning Research*, vol. 20, no. 112, pp. 1–49, 2019.
- [17] T. Haryanto, H. Suhartanto, and X. Lie, "Past, present, and future trend of gpu computing in deep learning on medical images," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 21–28, 2017.
- [18] L. Hou, N. J. Parmar, N. Shazeer, X. Song, Y. Li, and Y. Cheng, "High resolution medical image analysis with spatial partitioning," in *High Resolution Medical Image Analysis with Spatial Partitioning*, 2019.
- [19] W. Zhu, C. Zhao, W. Li, H. Roth, Z. Xu, and D. Xu, "Lamp: Large deep nets with automated model parallelism for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 374–384, Springer, 2020.
- [20] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, "Sample-efficient deep learning for covid-19 diagnosis based on ct scans," *medRxiv*, 2020.
- [21] D. Sierra-Sosa, B. Garcia-Zapirain, C. Castillo, I. Oleagordia, R. Nuño-Solinis, M. Urtaran-Laresgoiti, and A. Elmaghraby, "Scalable healthcare assessment for diabetic patients using deep learning on multiple gpus," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 10, pp. 5682–5689, 2019.
- [22] M. Usama, B. Ahmad, J. Wan, M. S. Hossain, M. F. Alhamid, and M. A. Hossain, "Deep feature learning for disease risk assessment based on convolutional neural network with intra-layer recurrent connection by using hospital big data," *IEEE Access*, vol. 6, pp. 67927–67939, 2018.
- [23] Y. Oyama, N. Maruyama, N. Dryden, E. McCarthy, P. Harrington, J. Balewski, S. Matsuoka, P. Nugent, and B. Van Essen, "The case for strong scaling in deep learning: Training large 3d cnns with hybrid parallelism," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1641–1652, 2021.
- [24] S. B. Akintoye, L. Han, X. Zhang, H. Chen, and D. Zhang, "A hybrid parallelization approach for distributed and scalable deep learning," 2021.
- [25] S. Guedria, N. De Palma, F. Renard, and N. Vuillerme, "R2d2: A scalable deep learning toolkit for medical imaging segmentation," *Software: Practice and Experience*, vol. 50, no. 10, pp. 1966–1985, 2020.
- [26] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyperparameter optimization," in *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), vol. 24, Curran Associates, Inc., 2011.
- [27] M. P. Ranjit, G. Ganapathy, K. Sridhar, and V. Arumugham, "Efficient deep learning hyperparameter tuning using cloud infrastructure: Intelligent distributed hyperparameter tuning with bayesian optimization in the cloud," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pp. 520–522, 2019.
- [28] C. Guo, L. Li, Y. Hu, and J. Yan, "A deep learning based fault diagnosis method with hyperparameter optimization by using parallel computing," *IEEE Access*, vol. 8, pp. 131248–131256, 2020.
- [29] K. Shankar, Y. Zhang, Y. Liu, L. Wu, and C.-H. Chen, "Hyperparameter tuning deep learning for diabetic retinopathy fundus image classification," *IEEE Access*, vol. 8, pp. 118164–118173, 2020.
- [30] R. J. Borgli, H. Kvale Stensland, M. A. Riegler, and P. Halvorsen, "Automatic hyperparameter optimization for transfer learning on medical image datasets using bayesian optimization," in *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, pp. 1–6, 2019.
- [31] V. S. Parvathy, S. Pothiraj, and J. Sampson, *Hyperparameter Optimization of Deep Neural Network in Multimodality Fused Medical Image Classification for Medical and Industrial IoT*, pp. 127–146. Cham: Springer, 2021.
- [32] P. Anbeek, K. L. Vincken, G. S. van Bochove, M. J. van Osch, and J. van der Grond, "Probabilistic segmentation of brain tissue in mr imaging," *NeuroImage*, vol. 27, no. 4, pp. 795–804, 2005.
- [33] H.-H. Chang, A. H. Zhuang, D. J. Valentino, and W.-C. Chu, "Performance measure characterization for evaluating neuroimage segmentation algorithms," *NeuroImage*, vol. 47, no. 1, pp. 122–135, 2009.
- [34] Google Brain Team, "tf.data: Tensorflow efficient input pipelines," October 2021. Available at <https://www.tensorflow.org/guide/data>.
- [35] Google Brain Team, "Tensorflow: Distributed training," October 2021. Available at https://www.tensorflow.org/guide/distributed_training.
- [36] Ray Project, "Ray.cluster overview," October 2021. Available at <https://docs.ray.io/en/latest/cluster/index.html>.
- [37] Ray Project, "Ray.SGD: Distributed TensorFlow," October 2021. Available at https://docs.ray.io/en/latest/raysgd/raysgd_tensorflow.html.
- [38] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.
- [39] Google Brain Team, "Tfrecord format," October 2021. Available at https://www.tensorflow.org/tutorials/load_data/tfrecord.
- [40] Google Brain Team, "Tensorflow profiler: Profile model performance," October 2021. Available at https://www.tensorflow.org/tensorboard/tensorboard_profiling_keras.
- [41] Google Brain Team, "Tensorboard: Tensorflow's visualization toolkit," October 2021. Available at <https://www.tensorflow.org/tensorboard>.
- [42] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019.
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
- [44] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017.
- [45] NVidia, "Using deepspeed and megatron to train megatron-turing nlg 530b," October 2021. Available at <https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model>.