Energy efficient deep learning for medical imaging: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Energy efficient deep learning for medical imaging

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

EEDL

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Deep learning has produced some of the most accurate and most versatile techniques for many applications in medical image computing and computer-assisted intervention. However, there are very few systematic rules to guide the design and training of deep learning methods. As a result, development of these methods typically involves a lot of guesswork, trial and error, and hyper-parameter search and fine-tuning. It has become very common to train tens or hundreds of models via a near-exhaustive search of the space of parameters that may influence the model performance. Moreover, the trend has been towards using larger and larger models and training them for longer hours. Researchers and practitioners spend a lot of electric energy in the hope of gaining small improvements in model performance. This trend has reached alarming proportions due to the growing popularity of deep learning models, the increasing availability of more powerful computational hardware that naturally consume more energy, and the gloomy outlook of the human-caused global warming. Therefore, there is great incentive for designing energy-efficient deep learning methods to reduce the carbon footprint of this indispensable technology. Specifically, we are in dire need of techniques that reduce the energy requirement of deep learning methods during both training and inference.

The goal of this challenge is to encourage the MICCAI community to innovate energy-efficient deep learning methods. The proposed challenge is timely because it encourages research and development that may address several very important concerns. The current practice in developing deep learning methods, which involves huge models, long training times, and massive architecture search and hyper-parameter fine-tuning, has several critical disadvantages:

1- It wastes much energy, as explained above.
2- It demands a great amount of time from the experts who develop these methods.
3- It makes it difficult to compare and contrast different methods in a fair manner. The seeming advantages of one method over another can be merely due to the extra time and energy spent on training, architecture search, and

hyper-parameter selection. This can also slow down the progress of the research community towards discovering new methods that are truly meritorious.
4- In competitions, challenges, and comparisons, it puts the teams with access to less computational resources at an unfair disadvantage and it biases the results of such competitions.
5- It limits the usability of deep learning methods for geographical areas where electric energy is less affordable, such as in many developing countries.

The proposed challenge can serve as a first step by the MICCAI community towards finding innovative solutions that can address these problems.

Challenge keywords

List the primary keywords that characterize the challenge.

Deep learning; energy efficiency; model development; developing countries; fairness; affordable AI.

Year

The challenge will take place in ...

2023

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

Importantly, in the revised proposal we are considering merging our challenge with a related MICCAI-2023 challenge proposal titled "Energy-efficient medical image processing", proposed by Dr. Michael Götz. This challenge will be based on LIDC-IDRI dataset, which is a large public dataset. Therefore, it can satisfactorily complement our heterogeneous MRI dataset that will be kept hidden. We anticipate that the synergistic combination of these two challenges will shed more light on the impact of pre-training on energy usage. We will encourage the participating teams to develop solutions that can address both challenges. We will analyze the effectiveness of different pre-training strategies on energy efficiency and performance across the two sub-challenges. We are highly optimistic that the joint challenge will lead to a deeper understanding of the factors that are involved in energy efficiency and a more comprehensive assessment of different methodological approaches.

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect to have at least 20-30 participating teams, hopefully closer to 50.

We have contacted many of our current and prior collaborators and colleagues in several institutions and several different countries. These include USA: Harvard University, Massachusetts Institute of Technology, University of

California Los Angeles, University of Southern California, University of California Merced, University of Florida, and Virginia Tech; Canada: The University of British Columbia, Simon Fraser University, University of Calgary, University of Manitoba, University of Saskatchewan, Laval University, India: Indraprastha Institute of Information Technology, UK: University of Leeds, Switzerland: University of Lausanne. They all have expressed great enthusiasm about the idea of this challenge and interest in participating in the challenge. Recent MICCAI challenges that have had a much more limited scope and more narrow audience have received tens of participating teams with complete submissions. For example, the Fetal Tissue Annotation (FeTA) Challenge in 2021 had 20 participating teams with successful submissions. We expect to attract a significantly larger number of teams.

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

All teams that successfully complete at least one of the two tasks in this challenge will be asked to provide a description of their methods. With input from all authors, the challenge organizers will write a manuscript to describe the results of the challenge and explain the main insights learned from the challenge. The manuscript will be submitted for publication in the journal Medical Image Analysis or in IEEE Transactions on Medical Imaging.

This will be done jointly with the other challenge mentioned above.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

There will be no on-site evaluations on the day of the event at MICCAI 2023.

The challenge evaluation will be performed off-site by the organizers. All evaluations will be completed well in advance of the conference date.

On the day of the event at MICCAI 2023, we will require a room to host the presentations by the selected teams. We will also have talks from invited speakers at the event. These will require a PC, projectors, loud speakers, and microphones.

TASK: Energy efficient methods for brain segmentation in MRI slices

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Please see the main abstract for the whole challenge above. The main abstract explains the motivation, background, and overarching goals of this challenge.

The overall goal of this challenge is to encourage the MICCAI community to innovate energy-efficient deep learning methods for medical imaging. It is unavoidable to restrict the challenge to certain task(s). To this end, we have decided to focus the challenge on the specific task of brain segmentation in MRI. Brain segmentation is an integral part of every computational pipeline in neuroimaging. We have extensive experience and a long track record of publications on automatic brain segmentation. Moreover, we have large and rich datasets of fetal MRI with manual brain segmentations that we have curated over the years.

Although focusing the challenge on an application is unavoidable, we aim to foster and promote methods that are more likely to translate successfully across applications. Therefore, the data used in this challenge will be heterogeneous:

1) The challenge data will include both 2D images (brain slices) and 3D images (brain volumes).
2) The challenge data will include multi-modal MRI: structural MRI, diffusion MRI, and functional MRI.

The challenge will be divided into two tasks based on data dimensionality. Task 1 will deal with 2D images, whereas Task 2 will be on 3D images. The participating teams will be allowed to take part in Task 1, Task 2, or both. However, their method is expected to work on all MRI modalities within the task. For example, a team that decides to take part in Task 1, should design a single method that will work on 2D structural, diffusion, as well as functional MRI data.

To further promote generalizability of the methods, and to discourage background (pre-challenge) energyconsuming
efforts that may optimize models for specific tasks, we will keep the training data mostly hidden. We will describe the characteristics of the training data to the participating teams, but provide only a few representative examples of the images and labels. Most of the training data will not be disclosed. The participating teams will submit their methods and training scripts to the challenge organizers, who will run the scripts to train and evaluate the methods.

****

This section explains Task 1 - Energy-efficient methods for brain segmentation in MRI slices

In this task, the participating teams are expected to develop methods to segment the brain in 2D images (slices).

The goal, as detailed in the following sub-sections, is to develop methods that:

1) Are trained to achieve high segmentation accuracy with as little electric energy as possible.
2) Segment test images (at inference time) with as little electric energy as possible.
3) Achieve high segmentation accuracy.

Keywords

List the primary keywords that characterize the task.

Deep learning, energy efficiency, brain segmentation, 2D segmentation, affordable AI.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Davood Karimi
Instructor at Harvard Medical School

Razieh Faghihpirayesh,
Ph.D. student at Northeastern University; Research assistant at Boston Children's Hospital

Hamza Kebiri
Ph.D. student at University of Lausanne

Arvind Balachandrasekaran
Research Fellow at Harvard Medical School

Clemente Velasco-Annis
Clinical Research Specialist at Boston Children's Hospital

Abdelhakim Ouaalam
Clinical Research Specialist at Boston Children's Hospital

Simon K. Warfield
Professor at Harvard Medical School

Ali Gholipour
Associate Professor at Harvard Medical School

The new organizing team members will include Dr. Michael Götz and Dr. Charlotte Debus, the organizers of the challenge proposal "Energy-efficient medical image processing".

b) Provide information on the primary contact person.

davood.karimi@gmail.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with fixed submission deadline.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

Challenge website will be announced once the challenge proposal is accepted.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

Additional points: Only fully-automatic segmentation methods will be considered.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Use of pre-trained models is permitted. Participants are allowed to pre-train their models on any data. However, in that case, the same pre-trained model should be trained (fine-tuned) and evaluated on all MRI modalities in this task.

The use of pre-trained models is a good strategy for reducing the energy requirement during training. Therefore, we encourage that. On the other hand, teams that have access to large amounts of similar data may spend much training time and energy to pre-train their model for the specific tasks in this challenge. This would defeat the purpose of the challenge. To avoid this situation: (1) We will keep the training data hidden, as mentioned above in the abstract for Task 1. (2) We demand that the same pre-trained model should be used on all MRI modalities.

We will require that the participating teams describe their pre-training methods in detail and disclose the data that they use for pre-training. This information will be used in interpreting the results of the challenge.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May not participate.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

At the moment, this challenge does not include an award. However, if the challenge is accepted, we will contact one or two corporations to support the challenge for a prize.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Participating teams that submit a method to this challenge will be considered fully committed to the challenge. We (the challenge organizers) reserve the right to publish the details of their methods and their performance results without identifying the participants' names or institutions.

We will ask for permission from the participating teams to also disclose their names and institutions in any presentations or publications. Participating teams can choose not to have their names/institutions publicly discloses. However, those teams will not be included in the official challenge event at MICCAI 2023 and will not be included in any publications that may result from this challenge.

We will announce the top 3 teams' names and institutions, if they grant us the permission to do so, at the challenge event at MICCAI 2023.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The top 3 teams can, each, have up to three authors. The remaining teams that successfully complete at least one of the two tasks, can each have up to two authors.

The participating teams will be permitted to publish their own methods and results at any venue. We will not apply an embargo. However, we apply two conditions: 1) They will not be allowed to include the methods and results of other participating teams until a paper describing the results of the challenge is published by the challenge organizers. 2) They should properly cite the references that will be specified by the challenge organizers on the challenge website.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participating teams will send their methods to the challenge organizers. Detailed submission instructions will be published on the challenge website.

The challenge organizers will run and evaluate the methods locally in their institution.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Each participating team will be allowed three pre-submissions for each of the two tasks. The challenge organizers will evaluate each pre-submissions and report, to the participating team, the details of the scores/metrics explained below.

After the (optional) pre-evaluations, each participating team will submit a final method that will be the basis of final evaluation and ranking.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Opening of the challenge website: One week after challenge proposal acceptance

Release of the training data: February 15, 2023
Registration deadline: March 15, 2023

Start of submission acceptance and pre-evaluation: April 1, 2023

End of submission acceptance: July 15, 2023

Top teams will be notified and speakers will be invited: August 15, 2023

Challenge event: October 2023

Manuscript preparation and submission: December 2023

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The study in which these data were acquired was reviewed and approved by IRB. The data was acquired in a study funded by the US National Institute of Health (NIH), and is shared based on the NIH data sharing policy. Only deidentified
data will be shared for this challenge. Since data is de-identified, it does not include patient health information and data sharing does not require IRB approval.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC ND.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The codes to evaluate the results and compute the metrics will become available on our lab's github repository. We will provide a link to the evaluation code on the challenge website once the challenge proposal is accepted. The evaluation code will be documented and the evaluation metrics will be clearly described to the participating teams.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

The participating teams will have to share with the challenge organizers either (1) their scripts for model training and testing, or (2) Dockerized versions of their methods.

They are also encouraged to release their codes publicly, but that will remain optional.

We will only accept Docker submissions.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

The challenge organizers have no conflicts of interest (financial or otherwise) to disclose.

All data used in this project were collected and labeled under projects that had been funded by the US National Institutes of Health.

Only the challenge organizers will have access to the test case labels. These labels will not be disclosed to the participating teams at any time.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

The goal of the challenge is to develop energy-efficient deep learning methods for medical image analysis. This specific task will include brain segmentation in 2D MR images.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

In utero MRI of fetuses between 18 and 38 gestational weeks.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

In utero MRI of approximately 500 fetuses scanned between 18 and 38 gestational weeks.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Structural, diffusion, and functional MRI.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Manual segmentation of the brain.

b) ... to the patient in general (e.g. sex, medical history).

None

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Fetal brain on in-utero MR images.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Fetal brain.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: Energy efficiency:
The main goal of this challenge is to develop energy-efficient deep learning methods. Therefore, the primary criterion used to rank the methods will be energy efficiency, which will be quantified as the amount of energy needed to achieve certain segmentation accuracy levels. Please see below for details.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Siemens Skyra, Prisma, and Trio MRI scanners. These were all 3T scanners.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Example structural imaging protocol: Multiple T2-weighted HASTE (Half-Fourier Single Shot Turbo Spin Echo) scans of the fetal brain in orthogonal planes were obtained with: TR= 1400–2000 ms, TE= 100–120 ms, 0.9–1.1 mm in-plane resolution, 2 mm slice thickness with no inter-slice space, acquisition matrix size= 256*204, 256*256, or 320*320 with 2- or 4-slice interleaved acquisition.

Example diffusion MRI protocol: Each session comprised 2–8 scans each along one of the orthogonal planes with respect to the fetal head. In each scan, 1 or 2 b= 0 s/mm^2 images, and 12 diffusion-sensitized images at b= 500 s/mm^2 were acquired. Acquisition parameters were: TR= 3000–4000 ms, TE= 60 ms, in-plane resolution= 2 mm, slice thickness= 2–4 mm.

Example functional MRI protocol: Axial fMRI volumes at 3mm resolution, 3mm slice thickness, parallel imaging acceleration factor of 2 with GRAPPA reconstruction, TR/TE of 2400-3000ms/40-70ms, number of measurements (volumes) = 80-250 per case.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

The scans were acquired at three different sites at Boston Children's Hospital.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Each training and test case is a 2D MR image (slice) of a fetal brain along with a manually-generated brain

segmentation mask.

b) State the total number of training, validation and test cases.

Structural (T2) MRI: 2000 training cases, 500 validation cases, and 500 test cases

Diffusion MRI: 600 training cases, 200 validation cases, and 200 test cases

Functional MRI: 600 training cases, 200 validation cases, and 200 test cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

In our experience with fetal brain segmentation in MRI slices, 1000-2000 training images with manual labels are sufficient to capture the heterogeneity and variability in the data. Likewise, 200-500 validation and test images with manual labels are quite sufficient for reliable assessment and comparison of methods in this application.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Our datasets have a remarkable richness in terms of variability in imaging center, scanner, and image quality.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Two experienced annotators, each with more than five years of experience in annotating neuroimaging data, manually annotated each of the images in the training, validation, and test datasets. Each image was annotated by one person.

A research fellow and a Ph.D. student independently inspected the quality of the annotations to ensure correctness and identify possible labeling errors.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The annotators did not require any specific instruction. They had prior training and extensive experience with annotating neuroimaging data and had performed similar annotations in the past. They used ITK-SNAP for all annotations.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The annotators had over five years of experience in annotating neuroimaging data. They had received training in neuroimaging annotation at Boston Children's Hospital. One of their main duties over the past five years has been to perform annotation and quality control on neuroimaging data, especially in MRI.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable. Only one annotator performed the annotation for each image. A research fellow and a Ph.D. student, independently, visually inspected and verified the correctness of each annotation.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

No pre-processing is applied on the imaging data.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The only possible source of error in brain segmentation is the ambiguity at the brain edges due to partial volume effect. All annotations have been visually inspected to ensure that they do not include any additional errors.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Energy consumption in terms of kWh of GPU energy usage.
Segmentation accuracy in terms of Dice Similarity Coefficient (DSC).
Segmentation error in terms of Hausdorff Distance (HD).
Segmentation error in terms of Average Surface Distance (ASD).

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The goal of the challenge is to design energy-efficient deep learning methods. More specifically, focusing on the task of brain segmentation in fetal MRI, the goal is to design methods that can achieve high segmentation accuracy with as little electric energy as possible.
We will assess energy consumption in terms of kWh of GPU energy usage.
We will assess segmentation performance in terms of DSC, HD, and ASD that are widely used metrics of segmentation accuracy/error.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

We will use three different criteria to score and rank the participating methods:
1- Energy usage for training.
2- Energy usage for inference.
3- Segmentation accuracy.

Therefore, teams that will complete this task will be ranked in three different ways. Details of these criteria are described below.

****
1- Energy usage for training

We will apply nnU-Net (Isensee et al. Nature methods 18.2 (2021): 203-211) to perform the segmentation task. The nnU-Net framework relies on hyper-parameter search and is considered to be the state of the art in medical image segmentation. We will allow nnU-Net to train to full convergence, without constraining the energy usage. We denote the DSC, HD, and ASD achieved by nnU-Net on the test set with DSC_ref, HD_ref, and ASD_ref. Additionally, we will record the amount of energy consumed by nnU-Net during training, which we denote with E_ref. This will be done by recording the GPU power (in kW) over time during training and computing the time integral of power to compute the energy.

We will consider the results achieved by nnU-Net as the baseline that the participating methods should match. We will record the amount of energy that the participating methods will have consumed when they will reach 90%, 95%, and 100% of the DSC achieved by nnU-Net. Consider the segmentation method for the participating team i, and let us denote the amount of energy that this method consumes to reach 90% of DSC_ref with E_i_90. We define E_i_95 and E_i_100 in a similar manner. We will compute the score for this method as follows:

Score_i = f(E_ref / E_i_90) + f(E_ref / E_i_95) + f(E_ref / E_i_100)

where f(x)= ReLU(x-1)

In other words, the participating teams will get a positive score if the amount of energy they need is less than that of nnU-Net. If the amount of energy consumed by a team is more than nnU-Net for any of the performance levels (90%, 95%, and 100%), the score will be zero for that level.

The participating team's method will be allowed to train until the amount of energy that it consumes reaches E_ref.

It is likely that some participating methods will not reach 95% or 100% of the DSC achieved by nnU-Net. Such methods will still be scored. If a method does not achieve 100% of DSC_ref, the third term in the score above will be zero. If it does not reach 95% of DSC_ref, the second term will also be zero. However, if a method cannot reach 90% of the DSC achieved by nnU-Net, it will be disqualified from this ranking.

For each team, the above score is computed separately for each of the three MRI modalities. The final score will be the average of the three scores.

****

2- Energy usage for inference

The participating teams' trained models will be run on the test set and the average amount of energy for segmenting a test image will be recorded. This is done separately for each of the three MRI modalities and then averaged to arrive at one score. Teams will be ranked based on the average amount of energy, in the reverse order. Similar to the first ranking approach above, here too we will only include those methods that reach at least 90% of DSC_ref.

****

3- Segmentation accuracy

Denote the DSC, HD, and ASD achieved on the test images by team i with DSC_i, HD_i, and ASD_i. Segmentation accuracy score for this team will be computed as follows:

Score_i = DSC_i / DSC_ref + HD_ref / HD_i + ASD_ref / ASD_i

The values of DSC, HD, and ASD in this equation will be the averages of these metrics over all test images in each modality. This score is computed separately for each of the three MRI modalities and then averaged to arrive at one score for each team. Participating teams will be ranked based on this score.

****

We will rank the participating teams based on each metric. We will arrive at a final ranking by averaging the ranks.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results will not be scored and will not be included in the rankings.

c) Justify why the described ranking scheme(s) was/were used.

Because the main interest of this challenge is energy-efficient methods, we allow for slightly lower segmentation performance than the state of the art methods trained with unconstrained energy usage. However, methods that significantly underperform will be disqualified from the ranking because energy saving should not come at the cost of a significant loss in performance.
The proposed ranking based on "Energy usage for training" is based on this rationale. Moreover, it encourages the teams to achieve 100% of the performance of the state of the art segmentation techniques.

Our second criterion, "Energy usage for inference", is also very relevant. This is because a model needs to be trained once, but may be applied on thousands or millions of images when deployed. Therefore, depending on the application, the energy consumed during deployment/inference may be even higher.

Our third criterion has to do with segmentation accuracy. This is not as important as the first two, which are by far more important. Nonetheless, since accurate brain segmentation is highly desired in neuroimaging, we will also separately rank the methods in terms of segmentation accuracy.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

In evaluating machine learning methods, training with different random weight initializations or cross-validation with different training and test datasets are used to generate a source of variability for statistical analysis. We will not have such variability in the results because a single split of our data into train/test subsets will be used to evaluate the methods. Moreover, each team will have one single score for each of the evaluation metrics. Therefore, the statistical analyses in this challenge will be limited.

Some of the analyses that will be performed to provide more insight regarding the goals of this challenge are listed below.

We will analyze the agreement in the ranking of the teams in the first and second rankings above. This will help determine whether methods that have energy-efficient training also have lower energy usage during inference. We can use the Friedman test for this analysis.

We will also analyze whether the first and second rankings will be different depending on the MRI modality. We can use the Friedman test for this analysis as well. Alternatively, we can carry out this analysis using the values (of energy consumption) instead of ranks. This can be done by building linear mixed models with and without random slopes and testing whether the model with the slope gives a significantly better fit.

We will further analyze the main factors of the methods that have influenced the rankings. Examples of these factors will include network architecture (fully-convolutional, attention-based, hybrid, etc.), pre-training, and model size (number of parameters) and depth (number of layers or modules).

We will analyze the segmentation results of different methods to detect the outliers. We will inspect the outliers to identify the causes of large segmentation errors. We will also assess whether the method rankings will change if outliers are removed.

b) Justify why the described statistical method(s) was/were used.

Please see 28a.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Please see 28a.

TASK: Energy efficient methods for brain segmentation in MRI volumes

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Please see above for the main abstract for the whole challenge and the abstract for Task 1.

The purpose of Task 2 is similar to Task 1, with the only difference that this task relates to segmentation of the brain in 3D images (MRI volumes). The goal of this task is to develop methods that:

1) Are trained to achieve high segmentation accuracy with as little electric energy as possible.
2) Segment test images (at inference time) with as little electric energy as possible.
3) Achieve high segmentation accuracy.

Keywords

List the primary keywords that characterize the task.

Deep learning, energy efficiency, brain segmentation, 3D segmentation, affordable AI.

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Same as Task 1

b) Provide information on the primary contact person.

davood.karimi@gmail.com

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with fixed submission deadline.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

Challenge website will be announced once the challenge proposal is accepted.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

Additional points: Only fully-automatic segmentation methods will be considered.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Use of pre-trained models is permitted. Participants are allowed to pre-train their models on any data. However, in that case, the same pre-trained model should be trained (fine-tuned) and evaluated on all MRI modalities in this task.

The use of pre-trained models is a good strategy for reducing the energy requirement during training. Therefore, we encourage that. On the other hand, teams that have access to large amounts of similar data may spend much training time and energy to pre-train their model for the specific tasks in this challenge. This would defeat the purpose of the challenge. To avoid this situation: (1) We will keep the training data hidden, as mentioned above in the abstract for Task 1. (2) We demand that the same pre-trained model should be used on all MRI modalities.

We will require that the participating teams describe their pre-training methods in detail and disclose the data that they use for pre-training. This information will be used in interpreting the results of the challenge.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May not participate.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

At the moment, this challenge does not include an award. However, if the challenge is accepted, we will contact one or two corporations to support the challenge for a prize.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Same as Task 1.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Same as Task 1.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Same as Task 1.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Same as Task 1.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Same as Task 1.

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Same as Task 1.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Same as Task 1.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Same as Task 1.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Same as Task 1.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

The goal of the challenge is to develop energy-efficient deep learning methods for medical image analysis. This specific task will include brain segmentation in 2D MR images.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

In utero MRI of fetuses between 18 and 38 gestational weeks.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

In utero MRI of approximately 500 fetuses scanned between 18 and 38 gestational weeks.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Structural, diffusion, and functional MRI.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Manual segmentation of the brain.

b) ... to the patient in general (e.g. sex, medical history).

None

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Fetal brain on in-utero MR images.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Fetal brain.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: Energy efficiency:
The main goal of this challenge is to develop energy-efficient deep learning methods. Therefore, the primary criterion used to rank the methods will be energy efficiency, which will be quantified as the amount of energy needed to achieve certain segmentation accuracy levels. Please see below for details.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Same as Task 1.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Same as Task 1.

Additionally, for structural (T2) MRI, we applied slice-to-volume registration and super-resolution reconstruction to compute volumetric brain images. This operation is common and it is needed to correct for inter-slice fetal head motion that can be excessive.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Same as Task 1.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Not applicable.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Each training and test case is a 2D MR image (slice) of a fetal brain along with a manually-generated brain segmentation mask.

b) State the total number of training, validation and test cases.

Structural (T2) MRI: 400 training cases, 50 validation cases, and 50 test cases

Diffusion MRI: 60 training cases, 30 validation cases, and 30 test cases

Functional MRI: 60 training cases, 30 validation cases, and 30 test cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Our experience has shown that 50-100 training images with manual labels are quite sufficient to develop accurate deep learning methods for fetal brain segmentation in MRI volumes. Also, 30-50 images with manual labels are sufficient for reliably evaluating different methods in this task.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Our data represent very rich variability in terms of imaging center, scanner, and image quality.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Same as Task 1.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Same as Task 1.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Same as Task 1.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Same as Task 1.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

For structural (T2) images, a validated slice-to-volume registration and super-resolution reconstruction technique was used to compute volumetric brain images from multi-stack slices.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Same as Task 1.

b) In an analogous manner, describe and quantify other relevant sources of error.

None.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Same as Task 1.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Same as Task 1.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Same as Task 1.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results will not be scored and will not be included in the rankings.

c) Justify why the described ranking scheme(s) was/were used.

Same as Task 1.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Same as Task 1.

b) Justify why the described statistical method(s) was/were used.

Same as Task 1.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Same as Task 1.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] Salehi, Seyed Sadegh Mohseni, Deniz Erdogmus, and Ali Gholipour. "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging." IEEE transactions on medical imaging 36.11 (2017): 2319-2330.

[2] Salehi, Seyed Sadegh Mohseni, et al. "Real-time automatic fetal brain extraction in fetal MRI by deep learning." 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018.

[3] Karimi, Davood, et al. "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis." Medical Image Analysis 65 (2020): 101759.

Further comments

Further comments from the organizers.

None