Energy-efficient Medical Image Processing: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Energy-efficient Medical Image Processing

Challenge acronym

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

E2MIP 2024

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.

To curtail and reduce the impact that climate change has on our socio-economic live, saving energy is key. Data centers in general and modern-day AI applications in particular are electricity super-users. Recent studies have attempted at estimating the carbon footprint of common large-scale AI applications, highlighting the unsustainable, environmentally questionable path of current AI research. Despite this research, reducing or even monitoring the energy consumption needed for computational approaches in medical imaging are still poorly investigated. We counteracted to this situation, i.e. the model development purely under the perspective of predictive performance while disregarding the accompanied environmental consequences, by the first round of the Energy-Efficient Medical Image Processing challenge during the last MICCAI. Given the ongoing importance of this topic, we aim to continue this important line of research. Especially given the focus of this years MICCAI on developing countries and the huge financial burden of high computing/ high energy solutions.

The goal of the challenge is to raise awareness for energy consumption of training and inference methods and foster the development of novel best-practice approaches and solutions to improve the energy efficiency of commonly used DL/ML/MIP models. This will hopefully increase the awareness of energy consumption in needed for medical image processing and lead to novel approaches that allow more efficient algorithms. In addition to this, we try to gather more information about the current situation w.r.t. energy-efficient computation on medical image processing, for example, the ratio of training and inference runs with an additional survey. Towards this end, the challenge will offer two pathways to develop energy-efficient medical image processing models:
* The original challenge will call for the submission of training and inference on a dedicated public dataset (the actual training/test-split is hidden to the participants) for three common tasks: segmentation, detection, and classification.
* To foster best practices and reporting of energy consumption in general AI model development, co-submission for inference and possibly training for other challenges will be offered. This was also done last year as a post-challenge analysis for a few challenges and we aim to build on the already established collaborations.

Each submission will be evaluated on the tier-2 Supercomputer HoreKa, located at the Karlsruhe Institute of Technology (KIT), Germany. HoreKa allows for precise measurements of whole compute node energy consumption per run via internal power sensors that are part of Lenovo's XClarity Controller (XCC) and can be read via IPMI. The whole workload energy consumption of submitted solutions will be measured for a full training run and inference on the hold-out test set. Submissions will be offered a dedicated amount of computing resources (1 full node, equipped with 4 NVIDIA A100 GPUS with each 48 GB of VRAM, connected via NVLINK, and Intel Xeon Platinum 8368 CPUs with a total of 76 cores) to run training and testing. For energy measurements, the participants have to submit their solutions before the on-site event. The final solution will then be calculated and evaluated in a period of two to three weeks before MICCAI and the final results reported during the on-site event. To allow the participants a prior estimation of the success of their approach, we will present guidelines and tools to measure the energy consumption on standard hardware during an envisioned initial workshop and on the challenge website.

We expect a trade-off between required energy consumption and achieved performance. To account for this, we will not report a winning approach but rather a Pareto front between achieved performance and energy consumption. A selection of high-performance approaches on the Pareto front will be given the chance to present their solution and approaches during the on-site event. The actual selection will depend on the number of submissions; however, the presentations will be selected w.r.t. interest for the audience, performance, completeness of pre-experiments, and originality.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Energy efficiency, environment, sustainability, green machine learning

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

No additional MICCAI Workshop planned.

A starting online workshop on energy-friendly MIP is planned for June with open submission. The topic will be energy-efficient medical deep learning. During this workshop, the challenge will be introduced and a set of best practice and known trick will be communicated to give all participants a head starts. Beside that international research groups will be given the opportunity to present their findings.

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.

10-15 based on the experience from previous challenges (For example AI-HERO). However, this comes with a high uncertainty.

Publication and future plans

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

It is planned to organize an official challenge proceeding in cooperation with an established scientific publisher where the participants describe their approaches and can publish preliminary results.

A joint publication with all participants and organizers is planned after the successful challenge event during MICCAI. This publication will contain the final results from the challenge and discuss the findings made. The first / last authors will come from the organizational teams. Each participating team can nominate authors of this publication. Depending on the number of participating teams, the number of authors per team might be limited, for example to two authors per team.

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.

All computations are carried out before the final challenge event. As a consequent, only the typical conference material like projector, poster space etc.. is required.

TASK 1: Lung Nodule Segmentation and Classification

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.

Lung nodule segmentation and classification is a well-studies and known task. Multiple solutions are published and the LIDC-IDRI dataset is a large, public dataset that serves as base for multiple groups.

Therefore, the focus will be on a set of important standard tasks for medical image processing, namely segmentation, detection and classification using already known, public dataset. This allows to build upon existing expertise with the actual task rather than learning how to solve a specific problem. Hopefully, the participants will be motivated by this to submit creative solutions, for example a combination of handcrafted features and simple classifiers could lead to a sufficient performance with low energy consumption. This goal will be also further strengthened by not ranking the participants but rather providing a clear indication of trade-offs. For this, we will be using a pareto front to "rank" the submissions.

The complete dataset is known. However, the participants must submit a training and inference script packed in a docker container to the challenge organizers, who will run the scripts to train and evaluate the methods on a unknown data split.

Keywords

List the primary keywords that characterize the task.

Energy consumption, Sustainability, Deep Learning, Segmentation, Detection, Classification, Green AI, NSCLC, Lung Nodule

ORGANIZATION

Organizers

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

Michael Götz, michael.goetz@uni-ulm.de
Davood Karimi Davood.Karimi@childrens.harvard.edu

b) Provide information on the primary contact person.

Michael Götz, michael.goetz@uni-ulm.de
Davood Karimi Davood.Karimi@childrens.harvard.edu

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

One time event with fixed submission deadline. (Possible extension is currently discussed)

Challenge venue and platform

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

N/A

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

The submissions must be run on specific hardware in order to assess the energy consumption for different steps. Thus, grand-challenges cannot be used. Instead, a combination of CMT, Synapse and github for paper, container and code submissions are envisioned.

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

www.e2mip.org

Participation policies

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

Fully automatic.

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.

We will rely on a public dataset, the LIDC-IDRI dataset. There are no restriction from our side.

The training is done on the provided hardware with no additional network access. So only the provided training data can be used. Public pretrained networks can be used as long as they are well-documented and not trained on
the available dataset. The private dataset will not be shared to ensure a fair comparison.

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.

Members of the organizational teams can submit own approaches. It will be indicated that these results are from the organizational teams both during the first presentation and any following publication of the challenge results.

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

We do not aim to nominate a winner. Instead we want to highlight the different approaches that are on the pareto front. We envision a scientific prize for selected, innovative solutions on the pareto front. Details are currently in discussion.

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.

The goal of this challenge is not the identification of a winner, but rather show the spectrum of the solutions and give an idea about possible trade-offs between runtime, performance, and energy consumption. No absolute winner will be reported, but a pareto front of the solutions will be shown. The approaches on the pareto front will be highlighted. Each submission team can decide if their names and reports will be directly linked to their result, however this is strongly encouraged.

All participants will be required to submit a paper describing their approach that will be published in the envisioned challenge proceedings and will be used to compose the joint publication.

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 workshop results will be published as a journal paper, with all participating teams as co-authors. The max. number of authors per team will be decided on once the final number of teams is fixed. An embargo time of 9 months is envisioned for publishing the results of the challenge to allow a high-impact publication.

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.

o Four objects must be submitted for a challenge contribution.
A) A docker container for the training and inference of the chosen task via the chosen submission platform. This will be executed on a local hardware with the unknown data / datasplit.
B) The code necessary for the generation of the docker file. Private sharing of the code is allowed, although we encourage publication. The code submission is required to prevent bias and cheating for example by using unmentioned pretrained models.
C) A survey to gather information about the development process, including questions about the approx. number of training runs and inference runs, the type of hardware etc.
D) A short paper describing the approach possibly including preliminary results.
o It is envisioned to have each solution run on one computational node (1 HPC compute node, equipped with 4 NVIDIA A100 GPUs (48GM VRAM), connected via NVLINK, and 76 intel Ice Lake processors) for approximately three days calculation time for both training and inference.

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.

We aim to provide solutions for preliminary testing and assessment of the resource demand of the developed solutions. However, there will be a restriction due to the ongoing demand for resources.

In addition, we aim to give the participants information about how they can evaluate the energy consumptionon on their own hardware in an envisioned starting workshop and/or on the challenge website.

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

o May: Planned to openly start the challenge by providing the description, having the corresponding website online, and communicating the details such as submission process, deadlines, data, etc.
o June: Workshop on energy-efficient medical deep learning Energy-efficient Medical Image Processing
o May to six weeks before MICCAI: Participants can register to contribute. Registration is necessary to be able to acquire the needed resources.
o Four to two weeks before MICCAI: Submission of the final solutions.
o Release date: During a half-day on-site challenge session, including the presentation of the most interesting approaches.

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).

Only public data will be used during the challenge, no additional ethical approve is needed.

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.

Additional comments:
Public data (LIDC-IDRI) will be used for mode one, so no restrictions from our side.

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 evaluation code will be made public in a github repository. However, due to the technical restrictions, the energy measurement cannot be directly replicated by the participants and may not be included. Instead, the participants will be provided with a hand-on for own testing.

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

Participating teams should make their code available to the organizers to avoid unfair pretraining. To allow own publishing, we allow a code-publishing embargo after which the code should be made public.

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.

No conflicts of interest.

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. The chosen open dataset focuses on Lung nodules.

Task category(ies)

State the task category(ies)

Examples:

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

Segmentation and Classification

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.

The dataset used focuses on clinical cohorts, which will be the focus both for this challenge as well as the final applications.

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

We plan to use the public LIDC-IDRI dataset due to the size (> 1000 samples), good available public documentation, the good annotation (up to four annotations per lesion) , the possibility to include multiple tasks and the lively research based on this dataset. More information are given in the original publication. doi: 10.7937/K9/TCIA.2015.LO9QL9SX

Imaging modality(ies)

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

CT

Context information

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

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

Tumor segmentation and rating of the tumors. For more details, please refer to the original publication of the dataset at doi: 10.7937/K9/TCIA.2015.LO9QL9SX

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

Please refer to the original publication of the dataset at doi: 10.7937/K9/TCIA.2015.LO9QL9SX

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.

Lung CT

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.

Lung nodules

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)).

(1) Minimize energy requirements during training and test for lung nodule segmentation
(2) Minimize energy requirements during training and test for lung nodule classification

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).

Please refer to the original description of the dataset at doi 10.1118/1.3528204.

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

Data originates from multiple centers, namely Weill Cornell Medical College, University of California, Los Angeles, University of Chicago, University of Iowa, and University of Michigan . For more details see publication 10.1118/1.3528204

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.

Data originates from multiple centers, namely Weill Cornell Medical College, University of California, Los Angeles, University of Chicago, University of Iowa, and University of Michigan . For more details see publication 10.1118/1.3528204

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).

Patients with suspicious lung nodules. For more details, please refer to the original publication of the dataset at doi 10.1118/1.3528204.

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.

One case is one 3d image. If multiple images are available for a patient, they will also be in the same split. (Patientwise data split)

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

We will use all available images within the LIDC-IDRI dataset (> 1000).

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

For the evaluation of the algorithm performance, a random 80%-20% patient wise split will be done. This split will be unknown to the participants to avoid extensive overfitting to the given split. We would prefer a complete separate dataset, however, this would conflict with the goal of using a open dataset to allow the usage of existing solutions. For this, we will rely on Task 2.

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.

The final split will be unknown to the participants to avoid extensive overfitting to the given split. A complete separate dataset would be good but would introduce additional computational challenges, diverting the focus from the original goal, e.g. energy efficiency.

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.

Each image was annoated by up to four clinical trained experts. For more details refer to the original publication: doi: 10.7937/K9/TCIA.2015.LO9QL9SX

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.

Please refer to the original publication of the dataset: doi. 10.7937/K9/TCIA.2015.LO9QL9SX

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.

Up to four clinical trained rater from different institutes created the annotations. For more details, see the original publication at doi: 10.7937/K9/TCIA.2015.LO9QL9SX.

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.

See 10.7937/K9/TCIA.2015.LO9QL9SX. In general, there are multiple annotations available for the nodules, from up to four different rater. In addition, each rater gave a rating on different properties like severity, speculation etc. which are going to be used for the classification task. For the segmentation task, the annotations will be generated using a fusion algorihtm like STAPLE.

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.

We will provide a standard implementation for each solution as starting point for the participants. Also a conversation into easy-to-use data formats. Beside that, no additional preprocessing will be applied to the 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.

Segmentation errors, errors in the classification labels due to manual estimations.

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 measure by hardware.
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).

Classification accuracy and ROC-Area under Curve (AUC)

All these metrics will be used to compute the rankings of the participating methods, as described below.

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

We selected the most-common metrics as major metrics for the test cases. This makes it easier to understand the trade-offs between energy consumption and performance.

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.

A pareto front can show the trade-off between performance as one axis and energy consumption as a second axis. We don't think that there is a unique metric to determine the optimal trade-off. We hope that the pareto front will further boost the development of innovative solutions. We aim to incorporate the uncertainties into the final visualization of the results, however this will depend on the number of submissions.

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.

No absolute ranking will be provided as we expect a trade-off between performance and energy consumption. To foster this approach, a pareto front of the results will be reported instead of a ranking.

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.

A confidence interval for the performance metrics will be quantified using bootstrapping. For energy consumption, a measure of uncertainty will be estimated from system properties and additional experiments. The uncertainty will be displayed using a pareto front and will be included in a tabular listing of the results.

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

Confidence intervals will allow for estimation of the accuracy and reliability of the findings. This seems to be of particular interest in this challenge because it affects the possible trade-offs. The uncertainty of the energy consumption should be measured per case, but this could lead to extensive resource requirements that cannot be fulfilled. From our experience, the energy consumption for larger jobs (runtime measured in days) is fairly accurate and stable.

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.

* Different settings w.r.t to training/inference ratios.
* Possibly the combination of algorithms with ensembling (performance gain vs. energy loss)

TASK 2: Fetal 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.

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 collected and annotated over the years.

Although focusing the challenge on an application is unavoidable, we aim to foster and promote energy-efficient methods that are more likely to translate successfully across applications. Therefore, the data used in this task will be heterogeneous. Specifically, the data will include multi-modal MRI: structural MRI, diffusion MRI, and functional MRI. The participating teams will submit one model that should segment the fetal brain in all three MRI constrasts. To further promote generalizability of the methods, and to discourage background (pre-challenge) energy-consuming 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 or Docker containers to the challenge organizers, who will run the scripts to train and evaluate the methods.

Keywords

List the primary keywords that characterize the task.

Energy efficiency, Deep Learning, Segmentation, Fetal brain, MRI

ORGANIZATION

Organizers

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

Michael Götz, michael.goetz@uni-ulm.de
Davood Karimi Davood.Karimi@childrens.harvard.edu

b) Provide information on the primary contact person.

Michael Götz, michael.goetz@uni-ulm.de
Davood Karimi Davood.Karimi@childrens.harvard.edu

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

One time event with fixed submission deadline. (Possible extension is currently discussed)

Challenge venue and platform

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

N/A

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

The submissions must be run on specific hardware in order to assess the energy consumption for different steps. Thus, grand-challenges cannot be used. Instead, a combination of CMT, Synapse and github for paper, container and code submissions are envisioned.

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

www.e2mip.org

Participation policies

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

Fully automatic.

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.

The dataset will be held in private.

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.

Use of pre-trained models is permitted. Participants are allowed to pre-train their models on any data. However, 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. (2) We demand that the same pre-trained model should be used on all MRI modalities. (3) We will require that all participating teams to disclose all data used for pre-training.

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

We do not aim to nominate a winner. Instead we want to highlight the different approaches that are on the pareto front. We envision a scientific prize for selected, innovative solutions on the pareto front. Details are currently in discussion.

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.

The goal of this challenge is not the identification of a winner, but rather show the spectrum of the solutions and give an idea about possible trade-offs between runtime, performance, and energy consumption. No absolute winner will be reported, but a pareto front of the solutions will be shown. The approaches on the pareto front will be highlighted. Each submission team can decide if their names and reports will be directly linked to their result, however this is strongly encouraged.

All participants will be required to submit a paper describing their approach that will be published in the envisioned challenge proceedings and will be used to compose the joint publication.

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 at 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 at 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 at 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 at 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).

Ethics approval has been obtained from the Institutional Review Board at Boston Children's Hospital.

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.

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 at Task 1.

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

Same at 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.

No conflicts of interest.

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 fetal 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 fetal 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)).

(1) Segment the fetal brain in MRI slices with high accuracy.
(2) Minimize energy requirements for model training and test.

The goal of this task is to develop energy-efficient deep learning methods. Therefore, the primary criterion used to assess and compare the methods will be energy efficiency, which will be quantified as the amount of energy needed to achieve certain segmentation accuracy levels.

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 between 2 and 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).

All these metrics will be used to assess and compare different methods.

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

The goal of the challenge is to develop energy-efficient deep learning methods for medical image analysis. The goal of this task, which focuses on fetal brain segmentation in MRI, 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.

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.

Link to the publication of LIDC-IDRI: https://aapm.onlinelibrary.wiley.com/doi/epdf/10.1118/1.3528204
Link to the public dataset:
https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

Further comments

Further comments from the organizers.

For task 1, we know that using a public dataset in a challenge comes with additional risks. However, we decided to do it anyways after long discussions as we wanted to keep the focus to the energy-consumption rather than the actual task. Nevertheless, we added counter measures to prevent an unfair or biased training, for example limiting the pre-training, the requirement to make the source code available and keeping the final training / test-split unknown to participants.