Quantification of Uncertainties in Biomedical Image Quantification: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Quantification of Uncertainties in Biomedical Image Quantification

Challenge acronym

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

QUBIQ

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.

A preliminary study on inter-observer variability of manual contour delineation of structures was carried out by L. Joskowicz et al. in 2019 and published in the journal 'European Radiology'. Its objective was to quantify the inter-observer variability of manual delineation of lesions and organ contours in CT images to establish a reference standard for volumetric measurements for clinical decision making and for the evaluation of automatic segmentation algorithms. It was observed that the variability in manual delineations for different structures and observers is large and spans a wide range across a variety of structures and pathologies. Two and even three observers may not be sufficient to establish the full range of potential variability of the outlines of the structures of interest. This variability, that is a property of the biological problem, the imaging modality, and the expert annotators, is – as of now - not sufficiently considered in the design of computerized algorithms for medical image quantification.

So far, uncertainties in predicted image segmentation are derived from general considerations of the statistical model, from resampling training data sets in ensemble approaches, or from systematic modifications of the predictive algorithm as in 'drop-out' procedures of deep learning procedures. At the same time, the definition of when the outline of an image structure to be quantified is 'uncertain' is a task- and data-dependent property of the quantification that can – and maybe has to – be directly inferred from human expert annotations. So far, there are no data sets available for evaluating the accuracy of probabilistic model predictions against such expert generated truth and there is no consensus on what procedures for uncertainty quantification return realistic estimates, and what procedures do not.

The purpose of the challenge is to benchmark algorithms returning uncertainty estimates (probability scores, variability regions, etc) of structures in medical imaging segmentation tasks. Specifically, the algorithmic output will be compared against uncertainties that human annotators attribute to the local delineation of various image

structures of diagnostic relevance, such as lesions or anatomical structures. Structures in several CT an MR image data sets have repeatedly been annotated by a group of experts to quantify the variability of boundary delineations. Tasks include the segmentation of lesions, such as brain tumors, lung tumors, liver tumors, brain hemorrhages, as well as anatomical structures, such as kidneys, and prenatal brains.

Challenge keywords

List the primary keywords that characterize the challenge.

Segmentation, uncertainty quantification, observer variability, probabilities

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

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 would expect 15-30 submissions. Together with invited speakers, organizers this will make about 30-50 participants.

Publication and future plans

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

We will make an arxiv paper available shortly after the challenge, serving as an initial reference for the data set. We anticipate the submission of a final journal article after the 2021 edition of the challenge, targeting MedIA or TMI. All participants with viable submissions (i.e., functional Docker containers) will become co-authors.

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.

We will use an online distribution / submission system (the same as for BRATS).

A standard seminar/lecture room will be sufficient for the workshop.

TASK: Quantifying segmentation uncertainties

SUMMARY

Keywords

List the primary keywords that characterize the task.

Segmentation, uncertainty quantification, observer variability, probabilities

ORGANIZATION

Organizers

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

Bjoern Menze, TUM
Leo Joskowicz, The Hebrew University of Jerusalem
Spyridon Bakas, University of Pennsylvania
Andras Jakab, University of Zürich
Ender Konukoglu, ETHZ
Anton Becker, MSKCC & Unispital Zürich
Christoph Berger, TUM

b) Provide information on the primary contact person.

Bjoern Menze, TUM - bjoern.menze@tum.de, http://home.in.tum.de/~menze/

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

The initial challenge will be held at MICCAI 2020. We intend to repeat the challenge at MICCAI 2021 to grow a community surrounding it, and – eventually – to broaden tasks and datasets.
Training data will be made available during the whole time, and after the challenge. There will be a live leaderboard with the public validation datasets that will also be used all year round. The test data used during the testing phase of the challenge will remain private and hidden, and challenge participants will have to submit Docker containers to obtain their ranking. This private test data will only be evaluated as part of the 2020 and 2021 MICCAI challenges.

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.

We will use a challenge platform at UPenn that is also used for the BRATS challenge (see https://www.cbica.upenn.edu/BraTS19 and https://ipp.cbica.upenn.edu/).

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

NA

Participation policies

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

Fully automatic.

Additional points: The participants will have to submit Docker containers with implementations of their algorithms. This will require fully automated methods.

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.

Participants can use private data, or modify the public challenge training data as they wish. They will have to disclose the use of additional or modified training datasets in their description of the algorithm, though. They will also have to detail on other training resources, in particular, the specifications of computing resources used, such as single GPU, or high performance computing (HPC) cluster with X number of nodes and Y number of GPUs.

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 organizers' research groups can participate, but are not eligible for awards.

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

The top three ranking methods will be publicly named and awarded diplomas. They are eligible for taking selfies with the organizers.

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.

All results will be made public, and will be published in the final challenge journal paper (targeting MedIA or IEEE TMI). If a participant wishes to retract their submission after results are made public, their contribution will be reported in an anonymized fashion. (The only reason for a valid and unreported retraction are obvious technical failures of the submitted Docker implementation of the algorithm.)

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).

All team members who contributed to the design of the algorithm will be named co-author in the final challenge paper (unless the results are retracted, see above). Every participant can publish their algorithms and results independently (in fact, they are encouraged to do so). They can only refer to quantitative results of other challenge participants after the challenge results are published officially as an arxiv paper draft which will be distributed after

the challenge in 2020 and submitted for publication after the challenge in 2021.

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.

The participants have to submit a Docker container following the same instructions as for the past BRATS challenges (https://github.com/BraTS/Instructions#docker-brats). Participants will have to submit one Docker file for each segmentation task / dataset, or a Docker implementation that can switch in between the tasks.

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.

The participants can use a 'leaderboard' to obtain quantitative scores on publicly available validation images. The leaderboard data are not used in the subsequent final testing phase, where the participants will obtain their ranking based on their Docker submission.

There are no limitations in how often participants can submit to the leaderboard. However, only one Docker submission will be accepted for the final evaluation/ranking on the private data.

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

Training and validation (public leaderboard) data will be made available before end of May.

Registration will be made possible after the challenge is announced.

Test data will be made available four weeks before MICCAI 2020.

Results will be released at MICCAI 2020. Top performing contributors will be informed in advance and invited to present their methods during the workshop.

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).

Not applicable, data has been made available to the public domain already.

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.

No code will be made available.

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

Submitted algorithm will be made publicly available in one centralized repository (Dockerhub / "Modelhub.AI"), but participants can opt out of this public release. Licenses will depend on the participants' preferences. Publicly available implementations will be specifically analyzed and highlighted in the arxiv and the final journal paper.

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 sponsoring is planned as of now. (See award section.)

Some of the organizers will have access to subsets or the full data at various time points. The organizers will not use test images or test labels during the training phase of submissions that their group will/may contribute to the challenge.

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

Research, Diagnosis, Quantification, Algorithmic development, General entertainment.

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.

The key objective is the evaluation of algorithmic models that predict label uncertainty in image segmentation tasks. As such, there will be several applications. We do not intend to benchmark diagnostic algorithms of one diagnostic application in particular, and there is not 'target cohort'.

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

Clinical datasets will be used. There are not specific inclusions or exclusion criteria (that would be of relevance to benchmarking probabilistic models). A majority of the data has been used in prior studies (see references given below).

Imaging modality(ies)

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

Magnetic resonance imaging, computed tomography imaging

Context information

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

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

No contextual clinical information will be made available.

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

No information on sex or medical history will be made available.

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.

There is no specific biomedical application that this challenge is targeting. Subsets used for benchmarking probabilistic delineations are as follows: CTs with liver tumors, lung tumors, kidney and brain hematomas. Tumor scans involved the use of contrast agents. MRI with brain tumors, prenatal brain scan, prostate. The tumor MRI data will comprise multi-parametric scans.

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.

The 'algorithmic target' is the quantification of uncertainties in the delineation of the structures of interest.

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)).

Find algorithms – as well as training or randomization methods of established segmentation methods – that are capable of reproducing the uncertainties of the experts' delineations.

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).

Magnetic resonance scanners and computer tomography scanners have been used. No detailed information about the type of scanner, center, etc. is available to us.

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

No comprehensive information on the imaging process is available to us.

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.

No comprehensive information on the center of data origin is available to us.

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).

No information on the persons who acquired the data is available to us.

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 patient dataset. It contains a 3D image volume (at least) with one 3D segmentation of the structure of interest, and several delineations of the structures of interest for one or several 2D slices in the same dataset. Slices are chosen to represent the variability of the structure of interest well, e.g., intersect with the center of the tumor.

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

Overall 490 CTs with annotated liver tumors (896 annotations, ~80 cases), with lung tumors (1,085 annotations, ~100 cases), with kidney contours (434 contours, ~ 40 cases) and brain hematomas (497 contours, ~ 45 cases). Tumor scans involved the use of contrast agents.
Overall 730 MRI with annotated brain tumors (6,000 annotations, ~ 600 cases), with contours of tissue boundaries of prenatal brain scan (500 annotations, ~50 cases), and with contours of the prostate and of prostate structures (800 annotations, ~80 cases). The tumor MRI data will comprise multi-parametric scans (from BRATS).

60 % of the data will describe the training set.
10 % of the data will describe the validation (public leaderboard) set.
30 % of the data will describe the testing (ranking) set.

Assignments to the three sets will be random.

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

Total number represent the maximal number of annotated datasets available to us at this point.
Proportions represent best practice in statistical learning.

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.

NA

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.

All tasks will have about 10 annotations for the same 2D reference slide. One full 3D annotations of the structures of interest is also available for each case and can be used during training. At this point, no measure will be taken to guarantee that annotators represent the most diverse set of decisions (e.g., annotators will not be asked to be 'maximally inclusive' or 'minimally inclusive' when annotating a given slices.)

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.

There are different tasks and datasets that all require different protocols. Full detail cannot be given here for reasons of brevity, but will be reported in the post-challenge paper.

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.

Different raters have annotated the various subsets following a specific annotation protocol per anatomical region of interest. All annotations have been done by (or supervised and revised by) experienced attending radiologists. Full details of each protocol cannot be provided at this point for reasons of brevity, but will be reported in the post-challenge paper, as already done for example for the BRATS challenge.

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.

No merging will take place.

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.

Datasets will be preprocessed separately. If multi-parametric data is used, the individual volumes are co-aligned to allow direct evaluation of the task in question, and avoid the inclusion of another factor of deviation from the main task. Measures to guarantee anonymization of all datasets are in place.

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.

Measuring the 'error' in the image annotation is the purpose of this challenge.

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

Technical error, human failure, lack of expertise.

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)

Dice Similarity Coefficient (DSC, Dice).

Additional metrics will also be calculated and reported during the challenge. Furthermore, alternative metrics will be assessed in the meta-analysis conducted by the challenge organizers, as an effort towards finding the optimal metrics for quantification and ranking.

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

Dice is a typical volumetric overlap score used in evaluating segmentation tasks, and hence challenge participants are familiar with DSC. The scores will be calculated for different thresholds to sample the probability distribution fully.

Dice will be calculated to assess a property of an algorithm after ground truth and prediction are binarized at various probability levels (that are 0.1, 0.3, 0.5, 0.7, 0.9). Dice scores for all thresholds will be averaged. This average will be used for the ranking.

Dice score have many shortcomings, in particular in the novel application domain of uncertainty-aware image quantification. To this end, additional metrics will be calculated and reported during the challenge. Furthermore, alternative metrics will be assessed in the meta-analysis conducted by the challenge organizers, as an effort towards finding the optimal metrics for UQ and ranking.

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.

Dice scores will be averaged across all tasks and all data sets. The participant performing best according to this average will be named the "winner" of the challenge.

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

Missing annotation will be treated as missing annotations, i.e., as submissions with all labels set to background.

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

No common standard has been established for ranking benchmark results across tasks and metrics. To this end, the "winner" of this challenge will be determined by a simple metric that has the benefit of being easy to implement and understood by challenge participants.

Alternative ranking schemes will be tested and their results will be reported to improve on this unfortunate situation and to offer better ranking schemes in the future.

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.

Information on the basic statistical analysis (score, ranking) is given above. Additional tests will be presented in the post-challenge meta-analysis journal paper.

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

See 27c.

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.

Experiments on fusing contributions and evaluating the ensemble performance / accuracy may be part of the analysis of the final journal paper, the same applies, for example, to search for bias or algorithmic variability.

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.

Full references will be given in the post-challenge paper.