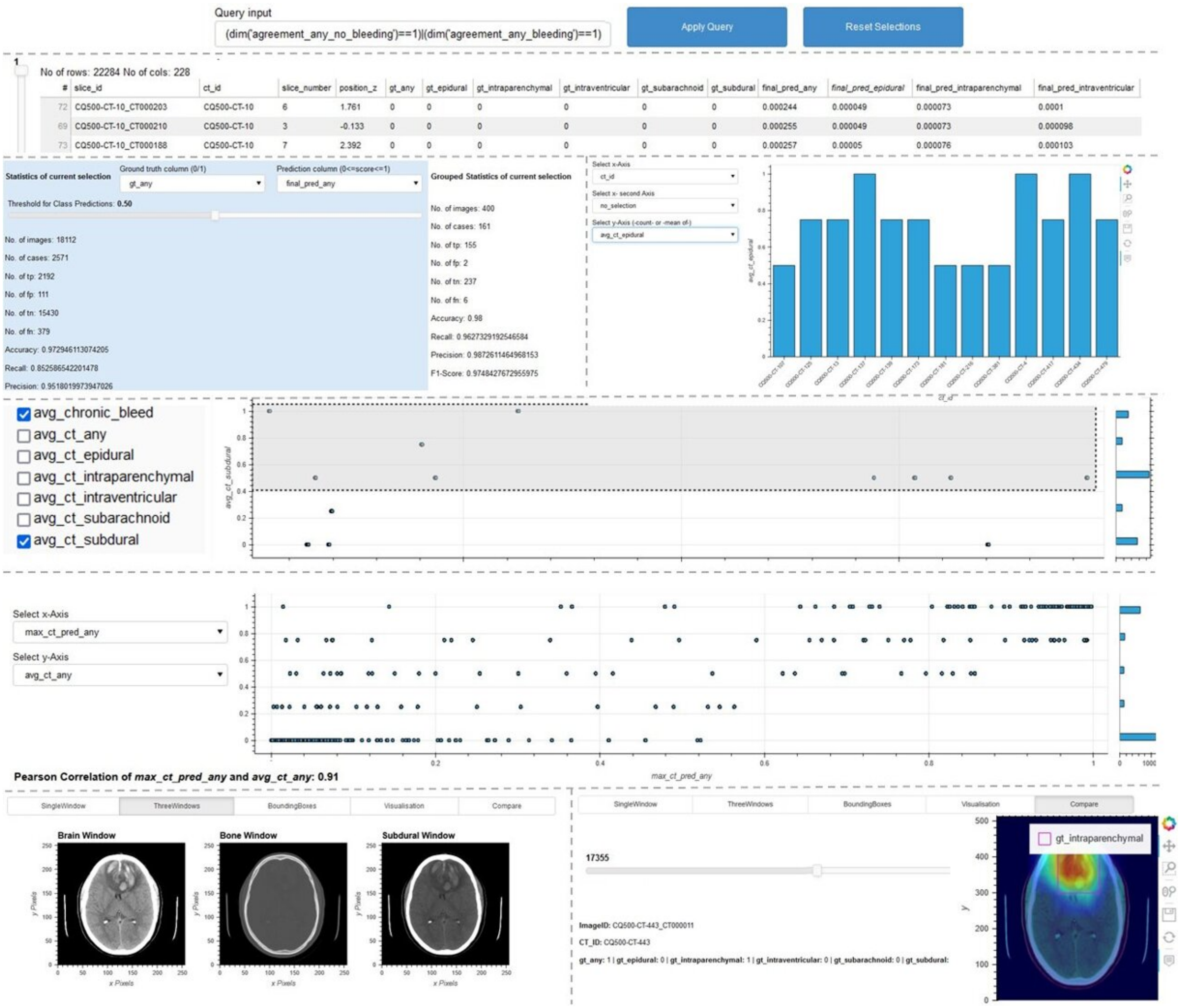
TechXplore

Designing trustworthy and transparent AI systems using assessment tools

April 3 2023



Query input

(dim('agreement_any_no_bleeding')==1)|(dim('agreement_any_bleeding')==1)

No of rows: 22284 No of cols: 228

#	slice_id	ct_id	slice_number	position_z	gt_any	gt_epidural	gt_intraparenchymal	gt_intraventricular	gt_subarachnoid	gt_subdural	final_pred_any	final_pred_epidural	final_pred_intraparenchymal	final_pred_intraventricular
72	CQ500-CT-10_CT000203	CQ500-CT-10	6	1.761	0	0	0	0	0	0	0.000244	0.000049	0.000073	0.0001
69	CQ500-CT-10_CT000210	CQ500-CT-10	3	-0.133	0	0	0	0	0	0	0.000255	0.000049	0.000073	0.000098
73	CQ500-CT-10_CT000188	CQ500-CT-10	7	2.392	0	0	0	0	0	0	0.000257	0.00005	0.000076	0.000103

Statistics of current selection

Ground truth column (0/1): gt_any

Prediction column (0<=score<=1): final_pred_any

Threshold for Class Predictions: **0.50**

No. of images: 18112

No. of cases: 2571

No. of tp: 2192

No. of fp: 111

No. of tn: 15430

No. of fn: 379

Accuracy: 0.972946113074205

Recall: 0.852586542201478

Precision: 0.9518019973947026

Grouped Statistics of current selection

No. of images: 400

No. of cases: 161

No. of tp: 155

No. of fp: 2

No. of tn: 237

No. of fn: 6

Accuracy: 0.98

Recall: 0.9627329192546584

Precision: 0.9872611464968153

F1-Score: 0.9748427672955975

Select x-Axis: ct_id

Select x- second Axis: no_selection

Select y-Axis (-count- or -mean of-): avg_ct_epidural

☑ avg_chronic_bleed
☐ avg_ct_any
☐ avg_ct_epidural
☐ avg_ct_intraparenchymal
☐ avg_ct_intraventricular
☐ avg_ct_subarachnoid
☑ avg_ct_subdural

Select x-Axis: max_ct_pred_any

Select y-Axis: avg_ct_any

Pearson Correlation of *max_ct_pred_any* and *avg_ct_any*: 0.91

17355

ImageID: CQ500-CT-443_CT000011

CT_ID: CQ500-CT-443

gt_any: 1 | **gt_epidural:** 0 | **gt_intraparenchymal:** 1 | **gt_intraventricular:** 0 | **gt_subarachnoid:** 0 | **gt_subdural:**

The ScrutinAI tool is used to detect errors in AI models or training data and to analyze the causes. In this example, an AI model for detecting abnormalities and diseases in CT images is being studied. Credit: Fraunhofer IAIS

The hype around ChatGPT has brought the topic of artificial intelligence and its impressive potential to the fore. At the same time, ensuring the quality and maintaining control of AI systems are becoming increasingly important—especially when these systems take on responsible tasks. After all, the chat-bot's results are based on huge amounts of text data from the internet.

That said, systems like ChatGPT only compute the most likely answer to a question and output it as a fact. Researchers from the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS will be showcasing various assessment tools and processes that can be used to systematically examine AI systems for weaknesses throughout their life cycle and safeguard against AI risks at the Hannover Messe 2023 from April 17 to 21 (at the joint Fraunhofer booth A12 in Hall 16).

The tools support developers and technical inspection authorities in systematically evaluating the quality of AI systems to ensure that they can be trusted.

As the media omnipresence of OpenAI's new AI application ChatGPT shows, artificial intelligence has reached an impressive level of maturity. The chatbot, trained with data and text from all over the internet, responds to questions with answers that are difficult if not impossible to distinguish from text written by humans. This makes the AI system worth considering for a wide range of tasks in companies, whether it's marketing tasks, automating how customer inquiries are handled or generating media content.

Assessment tools for peering into the black box

However, public discourse has also called for caution. The criticism is directed at, among other things, the lack of transparency, such as the sources the chatbot generates its answers from. In particular, the predictions are dependent on the quality of the input data.

"This shows how important it is to be able to systematically assess the quality of AI applications. Especially this is true in sensitive fields of application such as medical diagnostics, HR management, finance, the applications used by law enforcement agencies or safety-critical areas, where AI systems must deliver absolutely reliable results. The AI Act—the European draft for regulating AI systems—ranks these examples in the high-risk category and even requires mandatory assessment in these cases," says Dr. Maximilian Poretschkin, Head of safe AI and AI certification at Fraunhofer IAIS in Sankt Augustin, Germany.

"At this point, companies developing or deploying high-risk AI applications urgently need to address how they can ensure the quality of their applications."

Together with his team, he develops assessment tools and methods that examine and evaluate AI applications in terms of their reliability, fairness, robustness, transparency and data protection. The tools can be combined in a modular manner and are embedded in a software framework.

The development of prototypical assessment tools is supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia, among others, as part of the NRW ZERTIFIZIERTE KI (CERTIFIED AI) flagship project. The underlying assessment criteria are based on the AI Assessment Catalog, a structured practical guide published by Fraunhofer IAIS researchers in 2021.

Examining neural networks for vulnerabilities

The need for such assessment tools stems from the fact that AI applications often differ significantly from conventional software. The latter is programmed on the basis of rules, thereby enabling systematic testing of its functionality—i.e., whether the responses or outputs are correct in relation to the inputs. For AI applications, these procedures are in general not sufficient, especially if they are based on neural networks.

The ScrutinAI tool developed by Fraunhofer IAIS enables test personnel to systematically search for vulnerabilities in neural networks and thus assess the quality of AI applications. One specific example is an AI application that detects abnormalities and diseases in CT images. The question here is whether all types of abnormalities are detected equally well, or some better than others.

This analysis helps test personnel assess whether an AI application is suitable for its intended context of use. At the same time, developers can also benefit by being able to identify insufficiencies in their AI systems at an early stage and take appropriate improvement measures, such as enhancing the training data with specific examples.

It is conceivable that the tool could be used for many use cases. The above example could easily be replaced by an AI application that detects vulnerabilities and material defects in safety-critical components. In this case, too, it is important to establish whether all vulnerabilities are detected equally well or whether there are areas of the intended application domain for which the performance of the AI application is inadequate. "It's always about detecting insufficiencies in the neural network, albeit in different contexts," Poretschkin explains.

Assessing uncertainties

The uncertAInty method, developed by Fraunhofer IAIS and integrated into the framework, provides neural networks with a situation-dependent quality assessment function that they can use to evaluate their own certainty with respect to the prediction made.

"In highly automated AI decision-making, it is important to be able to assess how confident the AI is about the result it produces. To use a specific example, an autonomous vehicle must be able to reliably detect objects and people in its environment so that it can react to them appropriately. The uncertainty assessment helps in measuring how much you can trust the systems decision or whether certain fallback mechanisms need to be activated or if a human needs to make the final decision," says Poretschkin.

Therefore, the uncertAInty method constitutes an important building block for safeguarding AI applications so that they can be used in sensitive application contexts.

Comparing AI models

Lastly, the benchmarking tool is used to investigate which AI model is best suited for a particular task. "There's a glut of new AI applications that companies can integrate into their processes. Benchmarking helps them to make the right choice," says the researcher.

The tool has the functional capability, among other things, to measure the fairness of training data sets. This is crucial in the HR industry, for example, when it comes to AI applications that assist with the selection of new employees. In these situations, the AI application needs to be trained using balanced and statistically representative data sets to avoid disadvantaging groups of people and to ensure equal opportunities.

At the joint Fraunhofer booth A12 in Hall 16 at the Hannover Messe, the team from Fraunhofer IAIS will use an interactive demonstrator from the medical domain to show how the quality of an AI application can be systematically evaluated using the assessment tools. In addition, interested parties will find out how AI evaluation can be carried out in companies in practical terms.

More information: Conference: www.hannovermesse.de/en/

Provided by Fraunhofer-Gesellschaft

Citation: Designing trustworthy and transparent AI systems using assessment tools (2023, April 3) retrieved 17 April 2024 from https://techxplore.com/news/2023-04-trustworthy-transparent-ai-tools.html