Current Trends in AI and Cybersecurity

CSIT375/975 AI and Cybersecurity

Dr Wei Zong

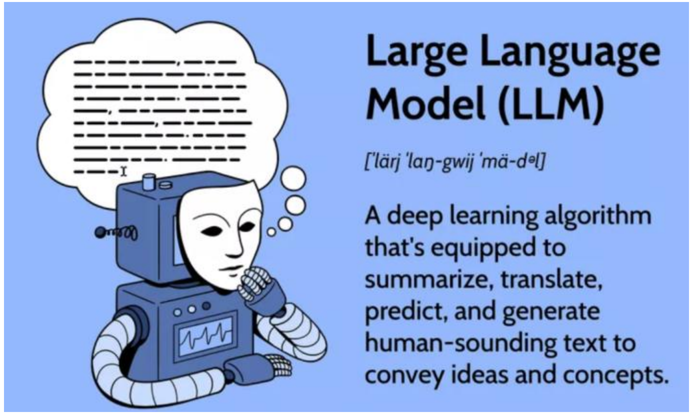
SCIT University of Wollongong

Disclaimer: The presentation materials come from various sources. For further information, check the references section

Outline

- Large language models (LLMs)
  - BERT
  - Detect LLM hallucinations.
- Explainable AI (XAI).
  - Grad-CAM
  - Mitigating spurious features.

What Is a Large Language Model (LLM)



Large Language Model (LLM)

['lärj 'laŋ-gwij 'mä-dᵊl]

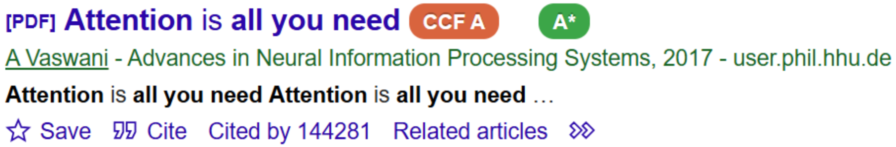
A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts.

- Large Language Models (LLMs)
  - LLM is a deep learning algorithm
    - ChatGPT, BERT, Llama, etc.
    - Summarize, translate, predict, and generate text to convey ideas and concepts.
      - Individual tasks have traditionally been solved by individual models created for each specific task.
    - LLMs rely on substantively large datasets to perform those functions.
      - ROOTS (1.6 TB): a diverse open-source dataset consisting of sub-datasets like Wikipedia and StackExchange for language modeling.

https://www.investopedia.com/large-language-model-7563532

LLM - BERT

- BERT
  - Short for Bidirectional Encoder Representations from Transformers.



[PDF] **Attention** is **all you need** CCF A A*

A Vaswani - Advances in Neural Information Processing Systems, 2017 - user.phil.hhu.de

Attention is **all you need Attention** is **all you need** …

    - **Transformers** was originally proposed for the language translation task.
    - Achieve excellent performance even in other fields, e.g., image recognition.
    - Developed in 2018 by researchers at Google.
  - BERT can be used on various language tasks, examples are:
    - **Sentiment Analysis**: determine how positive or negative a movie’s reviews are.
    - **Question answering**: help chatbots answer your questions.
    - **Text prediction**: predict your text when writing an email (Gmail).
    - **Text generation**: write an article about any topic with just a few sentence inputs.

Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

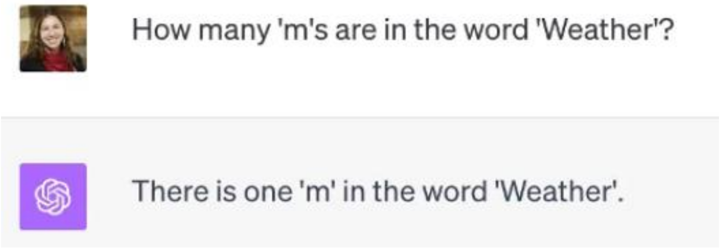
LLM - BERT

- Training BERT.
  - BERT is pre-trained on unlabeled data over 2 unsupervised tasks.
    - **Masked language model**
      - Mask some percentage, e.g., 15%, of the input words at random.
      - Then let the model predict those masked words.
    - **Next sentence prediction**
      - Choose two sentences A and B for each pretraining example.
        - 50% of the time B is the actual next sentence that follows A.
        - 50% of the time it is a random sentence from the corpus.
      - Let the model predict whether B follows A.
        - The final model achieves 97%-98% accuracy on next sentence prediction.
      - Help the model understands sentence relationships.
  - Fine-tune BERT using **labeled data** from the downstream tasks.
    - Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters.

Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Detecting Hallucination in LLM

- LLM Hallucination
  - LLM Hallucinations are the event when LLMs produce outputs that are coherent and grammatically correct but factually incorrect or nonsensical.
    - False or misleading information.



How many 'm's are in the word 'Weather'?

There is one 'm' in the word 'Weather'.

  - Potential reasons:
    - Limitations in training data.
    - Biases in the model.
  - Concern in fields that require high levels of accuracy
    - like healthcare, law, or engineering.

https://www.iguazio.com/glossary/llm-hallucination/

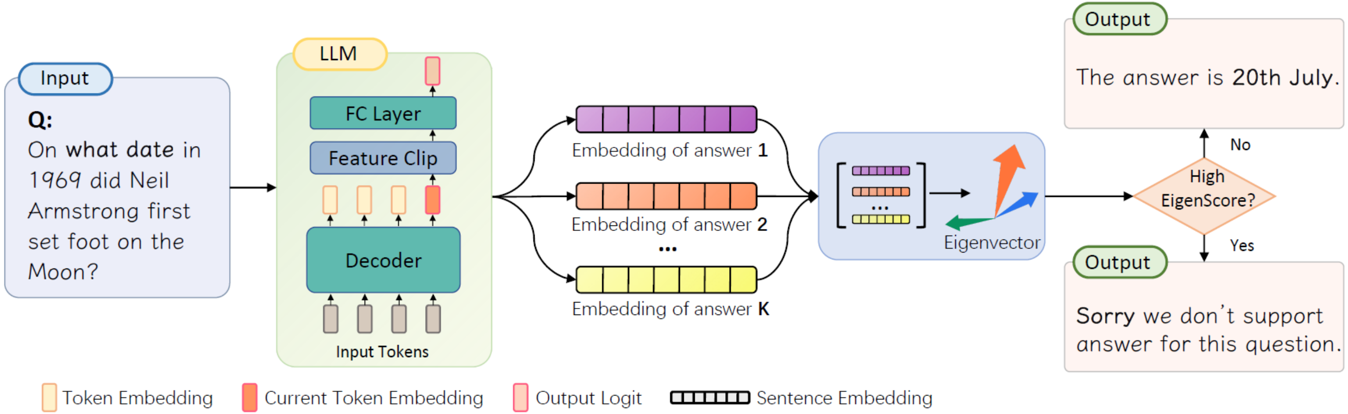
Detecting Hallucination in LLM

- LLM Hallucination Detection
  - Goal: detect and reject responses when hallucinations occur in LLMs.
  - Measure **uncertainty** has been proven effective
    - Prompting LLMs to **generate multiple responses to the same question**.
    - When a model is uncertain about its response, it generates hallucination context.
  - Use LLM internal states to measure uncertainty.
    - LLMs preserve the highly-concentrated semantic information of the entire sentence within their internal states.
    - Let LLMs generate multiple (e.g., 10) responses for one input.
      - Embeddings are extracted from LLMs internal layer.
        - Embeddings are numerical representations of text (like words, phrases, or entire documents) that capture their **semantic meaning**.
        - Embeddings are normally represented by vectors.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z. and Ye, J., 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. arXiv preprint arXiv:2402.03744.

Detecting Hallucination in LLM

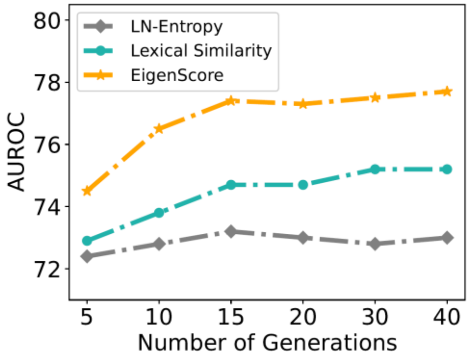
- LLM Hallucination Detection (continued)
  - The uncertainty degree can be measured by **EigenScore**.
    - Equivalent to the average logarithm of the eigenvalues calculated from the covariance matrix of sentence embeddings.
      - Technical details are not covered in this subject.
    - Small EigenScore -> the sentence embeddings are highly correlated.
    - Large EigenScore -> sentences contain diverse semantics.



Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z. and Ye, J., 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. arXiv preprint arXiv:2402.03744.

Detecting Hallucination in LLM

- LLM Hallucination Detection (continued)
  - Evaluation metric: AUROC
    - Short for **the area under the receiver operator characteristic curve**.
    - It is a popular metric to evaluate the quality of a binary classifier.
      - An AUROC score of 1 is a perfect score and an AUROC score of 0.5 corresponds to random guessing.
  - Performance of LLaMA-7B on Natural Questions dataset.
    - EigenScore consistently outperforms previous methods (i.e., LN-Entropy and Lexical Similarity) by a large margin for different number of generated responses.



Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z. and Ye, J., 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. arXiv preprint arXiv:2402.03744.

Explainable AI (XAI)

- What is Explainable Artificial Intelligence (XAI)?
  - XAI is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.
  - **Explainability** provides insight into the AI's decision to the end-user in order to build trust that the AI is making correct and non-biased decisions based on facts.
    - This is different from **interpretability**.
      - Interpretability enables developers to delve into the model' decision-making process, boosting their confidence in understanding where the model gets its results.
    - Mechanism of human brains cannot be fully **interpreted** yet, but human behaviors can be **explained**.
      - Drink when we are thirsty.
      - Eat when we are hungry.
      - Stop the car because of the red traffic light (and don't want huge fines).
      - Etc.

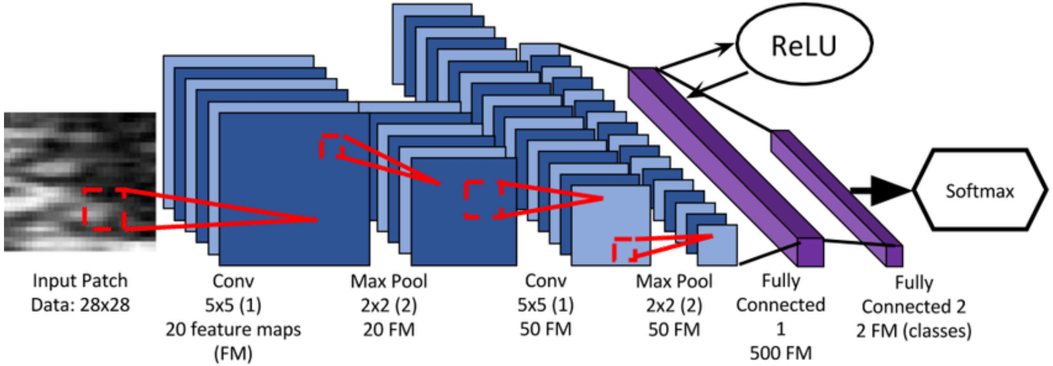
Grad-CAM

- Goal
  - A class-discriminative visualization technique.
    - Generate visual explanations from any **CNN-based** network
      - New techniques are being invented without dependence on CNNs.
      - Grad-CAM is a must-read study.
    - Do not require architectural changes or re-training of the target model.
  - Provide insight into failures of CNNs.
    - Able to show that seemingly unreasonable predictions have reasonable explanations.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Grad-CAM

- Approach
  - Key idea: Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to understand the importance of each neuron for a decision of interest.
    - The last convolutional layer typically learns high-level semantics.



https://www.researchgate.net/figure/The-overall-LeNet-architecture-The-numbers-at-the-convolution-and-pooling-layers_fig2_318972455

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

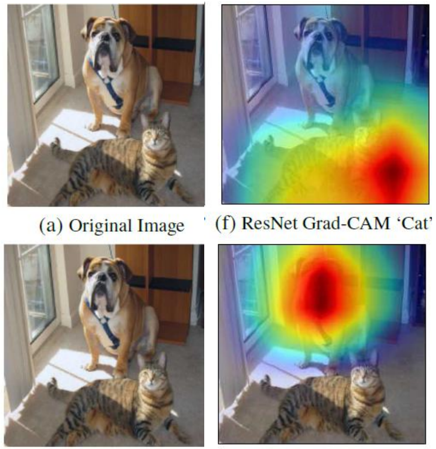
Grad-CAM

- Approach (continued)
  - Formula
    - Obtain the importance weight for a feature map: $$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$
      - y^c denotes the logit for class c.
      - A^k is the k^{th} feature map.
      - Z is the total number of elements, used for averaging.
    - Obtain visualization via a weighted combination: $$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$
      - ReLU is applied because visualization is only interested in the features that have a positive influence on the class of interest
        - i.e. features whose intensity should be increased in order to increase y^c.
        - Negative features are likely to belong to other categories in the image.
        - Without this ReLU, localization maps sometimes highlight more than just the desired class and achieve lower localization performance.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Grad-CAM

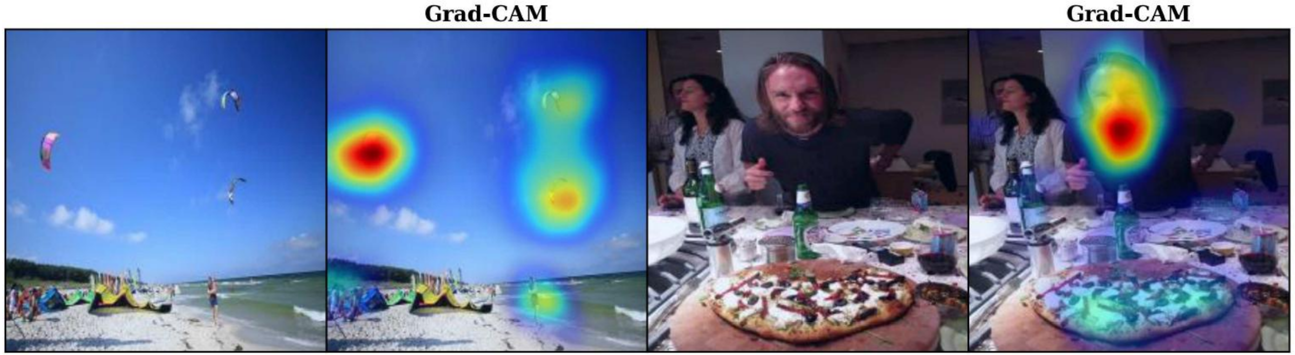
- Results on image classification



(a) Original Image (f) ResNet Grad-CAM ‘Cat’

(g) Original Image (l)ResNet Grad-CAM ‘Dog’

- Results on image captioning

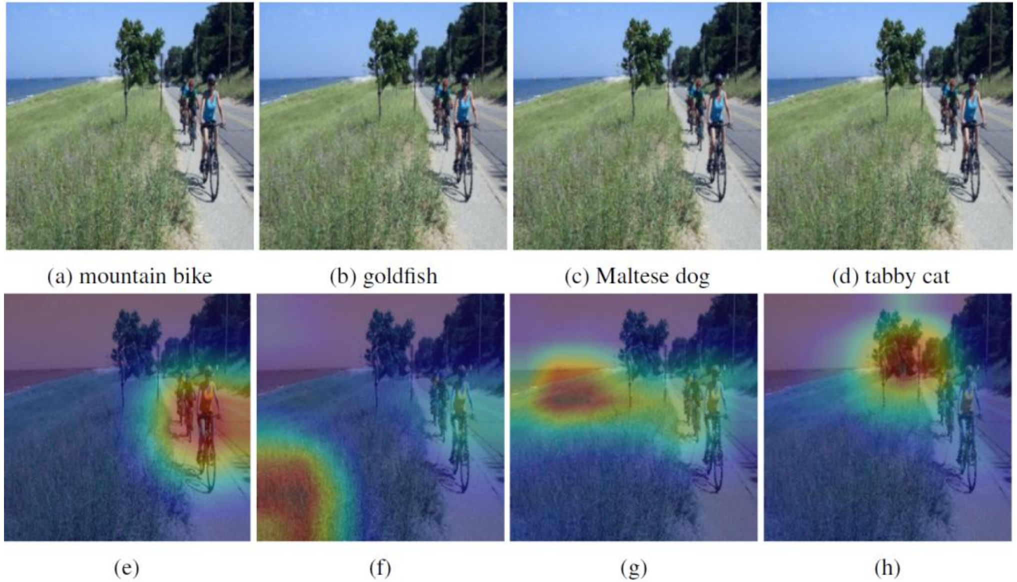


A group of people flying kites on a beach A man is sitting at a table with a pizza

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Grad-CAM

- Results on adversarial examples

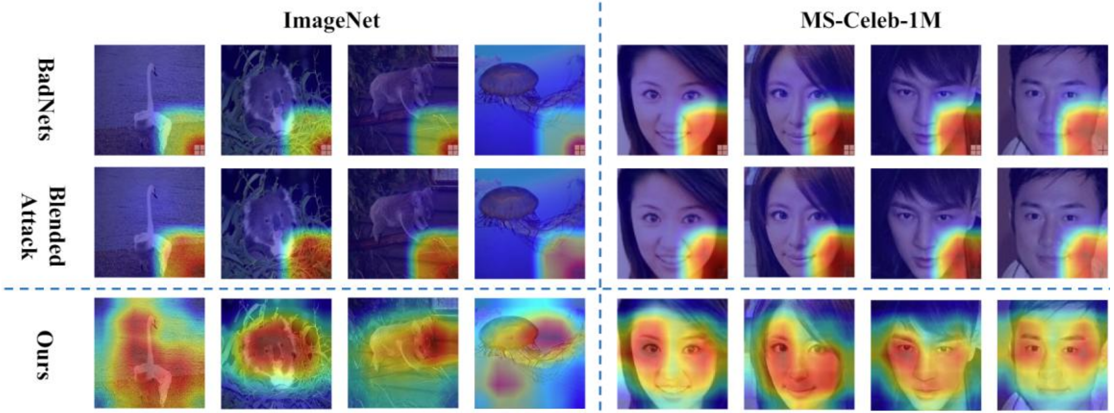


  - CAM visualization (a specialized version of Grad-CAM).
    - (a) the original image
    - (b)-(d) are adversarial examples targeting different classes.
    - Row 2 shows the visualization for the corresponding images above.

Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M. and Song, D., 2018. Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612.

Grad-CAM

- Results on backdoor attacks



- The Grad-CAM of poisoned samples generated by different attacks.
  - Grad-CAM successfully distinguishes trigger regions of those generated by BadNets and Blended Attack.
  - However, Grad-CAM cannot identify Invisible Backdoor Attack (SSBA).
    - Denoted as “Ours”.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

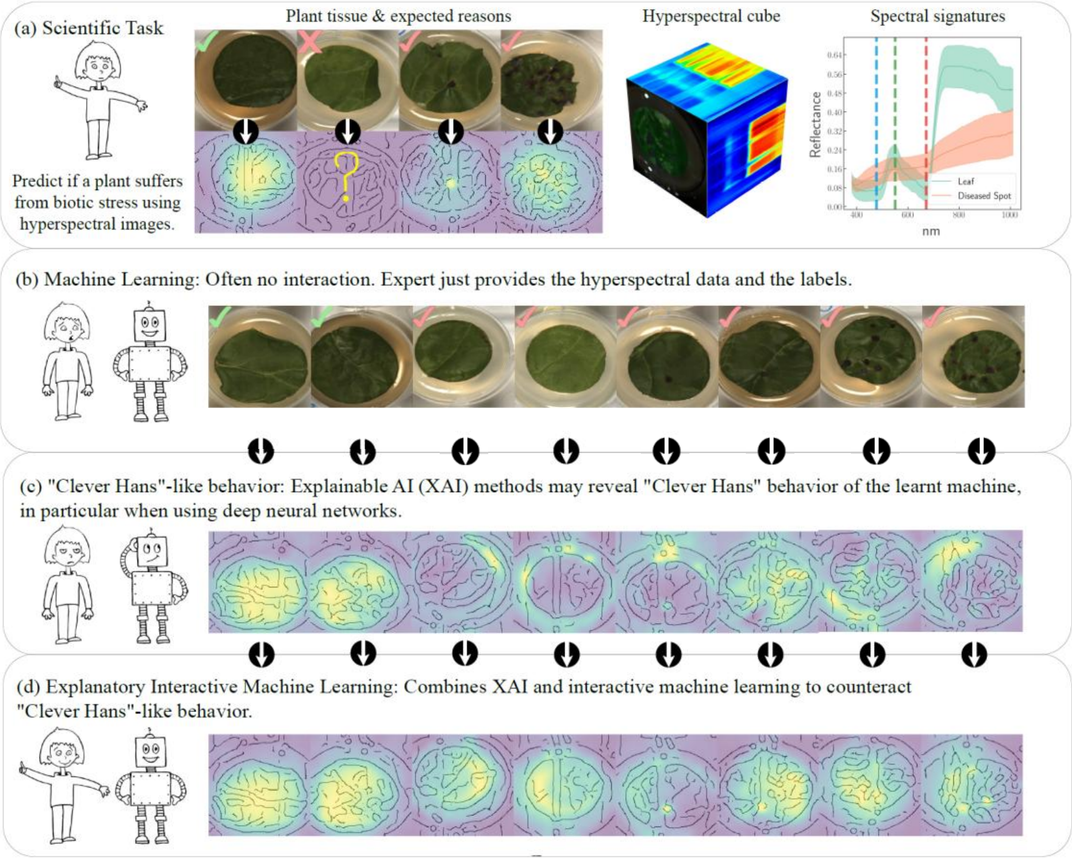
Making DNNs right for the right scientific reasons

- Problem setup
  - A plant phenotyping team is attempting to characterize crop resistance to plant pathogens.
    - The plant physiologist records a large amount of hyperspectral imaging data.
    - A machine learning expert is asked to apply deep learning to the data analysis.
      - The resulting predictive accuracy is very high.
      - The plant physiologist, however, remains skeptical: the results are "too good, to be true".
    - Using XAI techniques, they discover that the trained deep model uses clues within the data that do not relate to the biological problem at hand
      - Spurious features.
    - The machines have learned the right predictions for the wrong reasons and can therefore not be trusted.
  - Question: can we correct the model, towards making the right predictions for the right reasons?

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

- Human users revise learning machines towards trustworthy decisions.



(a) Scientific Task

Predict if a plant suffers from biotic stress using hyperspectral images.

(b) Machine Learning: Often no interaction. Expert just provides the hyperspectral data and the labels.

(c) "Clever Hans"-like behavior: Explainable AI (XAI) methods may reveal "Clever Hans" behavior of the learnt machine, in particular when using deep neural networks.

(d) Explanatory Interactive Machine Learning: Combines XAI and interactive machine learning to counteract "Clever Hans"-like behavior.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

- Correcting the model by a user.
  - Right for the right reasons
    - The prediction and the explanation are both correct.
    - No feedback from the user is requested.
  - Wrong for the wrong reasons:
    - The prediction is wrong.
    - The user is only asked to provide the correct label.
  - Right for the wrong reasons
    - The prediction is right, but the explanation is wrong.
    - Correct the explanation via counterexamples.
      - Identify areas of incorrect explanations.
      - Add modified copies of input into the training set.
        - The identified areas are either randomized, changed to an alternative value, or substituted with the value of corresponding component appearing in other training examples of the same class.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

- Alternatively, minimize gradients w.r.t features maps in irrelevant areas.

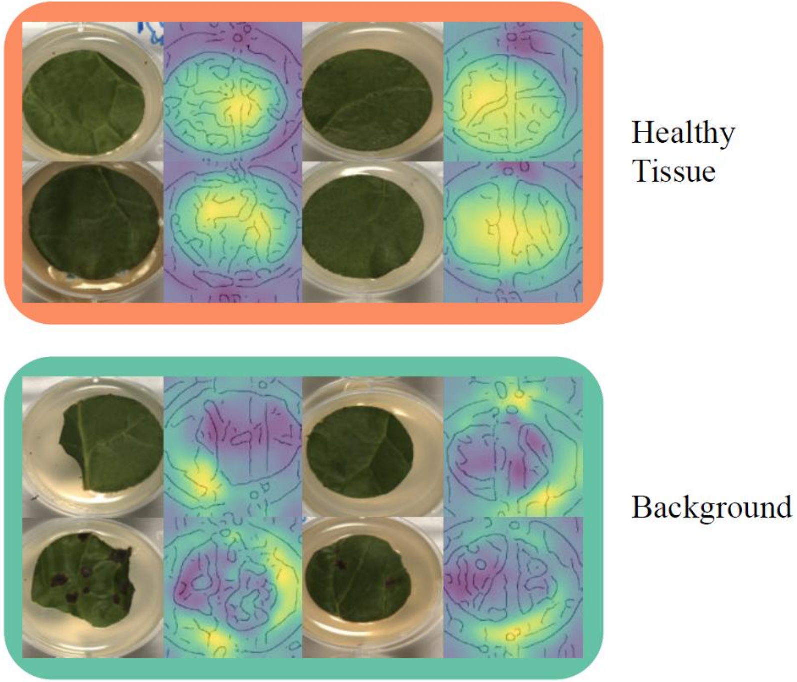
$$L(\theta,\ X,\ y,\ A) = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} -c_k y_{nk} \log(\hat{y}_{nk})}_{\text{Right answers}} + \underbrace{\lambda_1 \sum_{n=1}^{N}\sum_{d=1}^{D}\left(A_{nd}\frac{\delta}{\delta h_{nd}}\sum_{k=1}^{K} c_k \log(\hat{y}_{nk})\right)^2}_{\text{Right reasons}} + \underbrace{\lambda_2 \sum_i \theta_i^2}_{\text{Weight regularization}}$$

  - Right for the right reasons (RRR) loss.
  - The **first** and **third** terms are the traditional cross entropy loss.
    - $\hat{y}$ denotes the model prediction.
    - c is a rescaling weight given to each class of the unbalanced dataset.
  - The **second term** discourages the input gradient from being large in irrelevant regions for all labels.
    - Irrelevant regions are marked by a binary mask A.
    - h denotes the final convolutional layer.
      - Grad-CAM also uses this layer.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

- The deep neural network might be right for the wrong reasons.
  - Use RBG and hyperspectral images achieve 89% and 99% accuracy, respectively.
    - Models are trained with the cross-entropy loss.
    - Nearly perfect predictive performance is rather suspicious since plant phenotyping is a rather difficult task.
  - The network may focus on the nutritional solution for classification.
    - A biologist would consider this as cheating rather than valid problem-solving behavior.

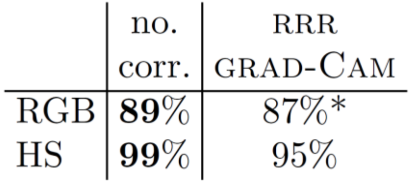


Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

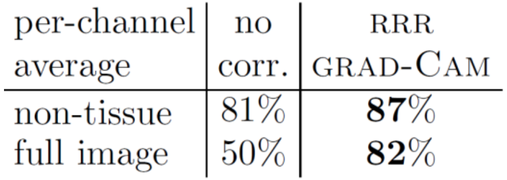
- Correct models based on GRAD-Cam
  - Evaluating the default and revised models on a test dataset without spurious background features.
    - The spurious background features were set to either the per-channel average of the non-tissue regions or the full image of the training samples.
    - Accuracy of the normally trained model decreases.
    - Accuracy of the corrected model shows generalization.
      - The corrected model focuses on image regions lying only on the tissue, regardless of the underlying class.

no corr.: without corrections; RRR: Right for the Right Reason



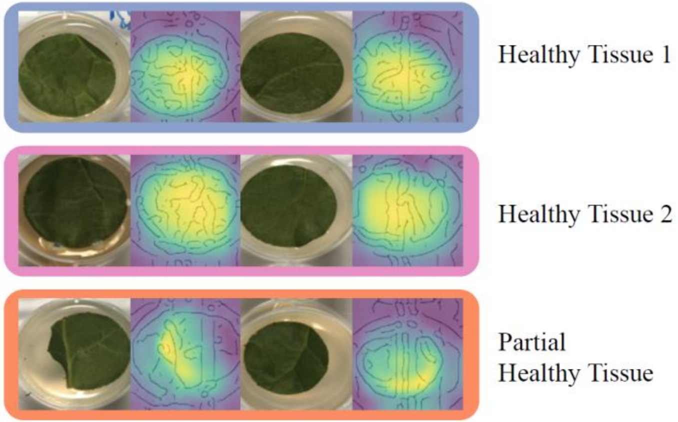
	no. corr.	RRR GRAD-CAM
RGB	**89%**	87%*
HS	**99%**	95%

Scientific Dataset



per-channel average	no corr.	RRR GRAD-CAM
non-tissue	81%	**87%**
full image	50%	**82%**

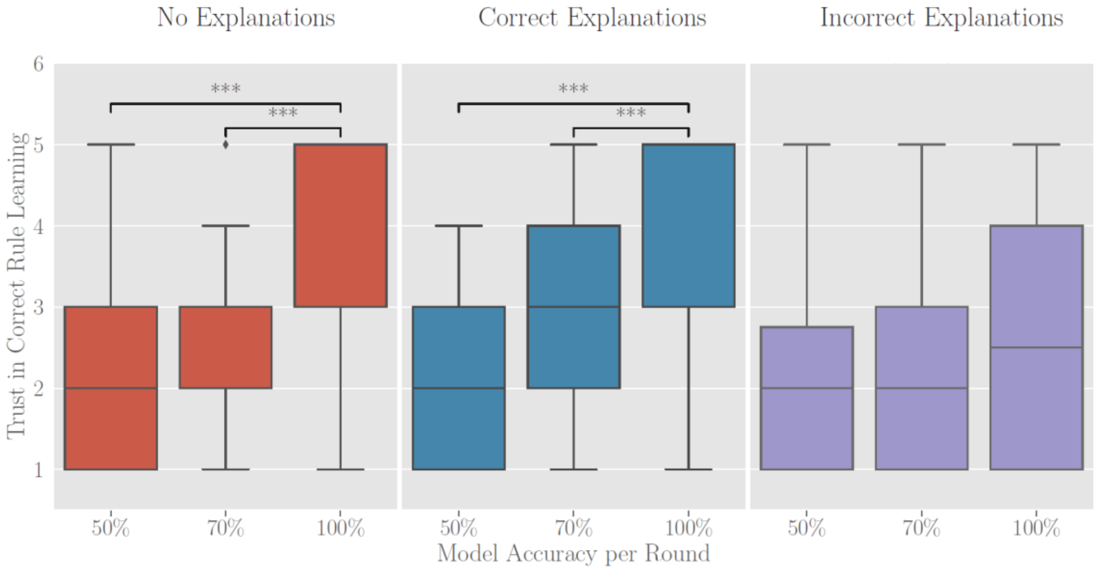
HS Scientific Dataset
Without spurious features



Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Making DNNs right for the right scientific reasons

- Trust in AI

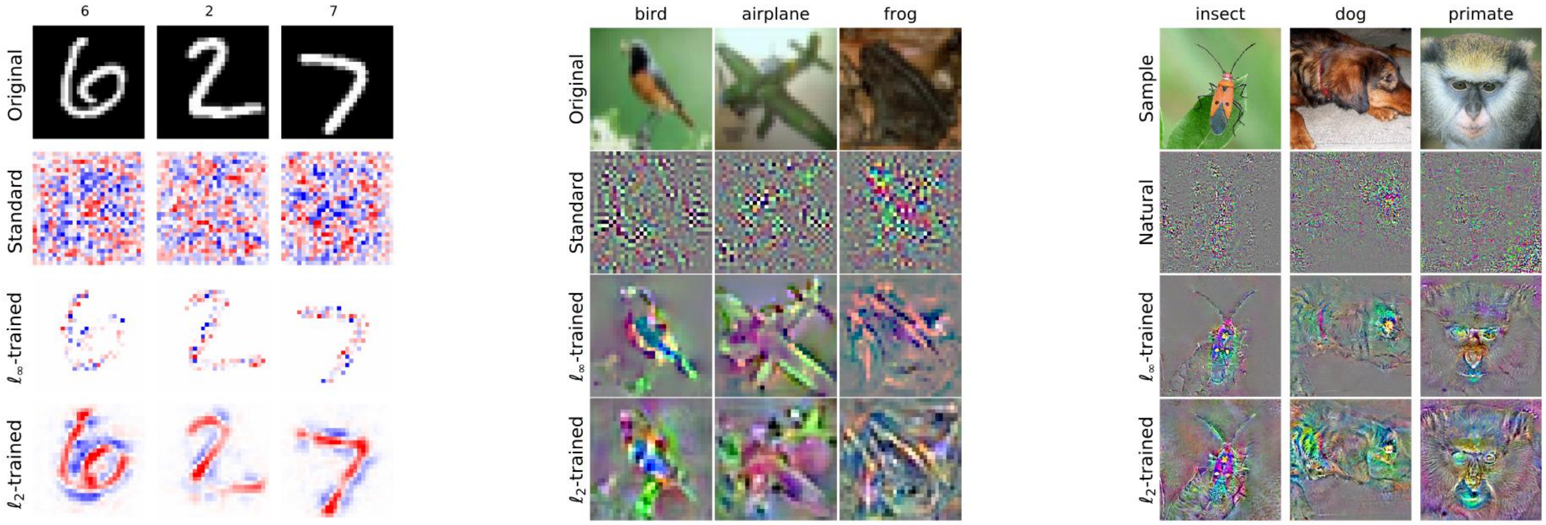


- 106 participants were asked to rate “I trust that the AI has learned the correct rule for classifying such images.”
  - Without explanations, people trust highly accurate machines.
  - Correct explanations can increase the trust when models’ performance is mediocre.
  - people do not forgive wrong explanations even if the predictions are correct.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.

Recall Adversarial Robustness

- Eliminate shortcut learning.
  - Empirical evidence shows that adversarial robustness is somehow related to human perception.
    - Visualization of the loss gradient with respect to input pixels for adversarially trained networks.



(a) MNIST (b) CIFAR-10 (c) Restricted ImageNet

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A., 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.

References

- Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z. and Ye, J., 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. arXiv preprint arXiv:2402.03744.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
- Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K. and Kersting, K., 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8), pp.476-486.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A., 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.