



VINUNIVERSITY



SOICT

Explainable AI: How humans can trust Artificial Intelligence?

Hieu Pham, Ph.D.
Research Fellow, VinUni-Illinois Smart Health Center,
VinUniversity

Hanoi, Nov. 13th, 2021

What is the problem?

- Artificial intelligence and machine learning (AI/ML) systems have exceeded human performance in nearly every application where they have been tried.
- AI systems can make mistakes, and human users will not trust their decisions without explanation.

→ **This explainability requirement lead a new area of AI research, know as Explainable AI (XAI).**

What is explainable AI?

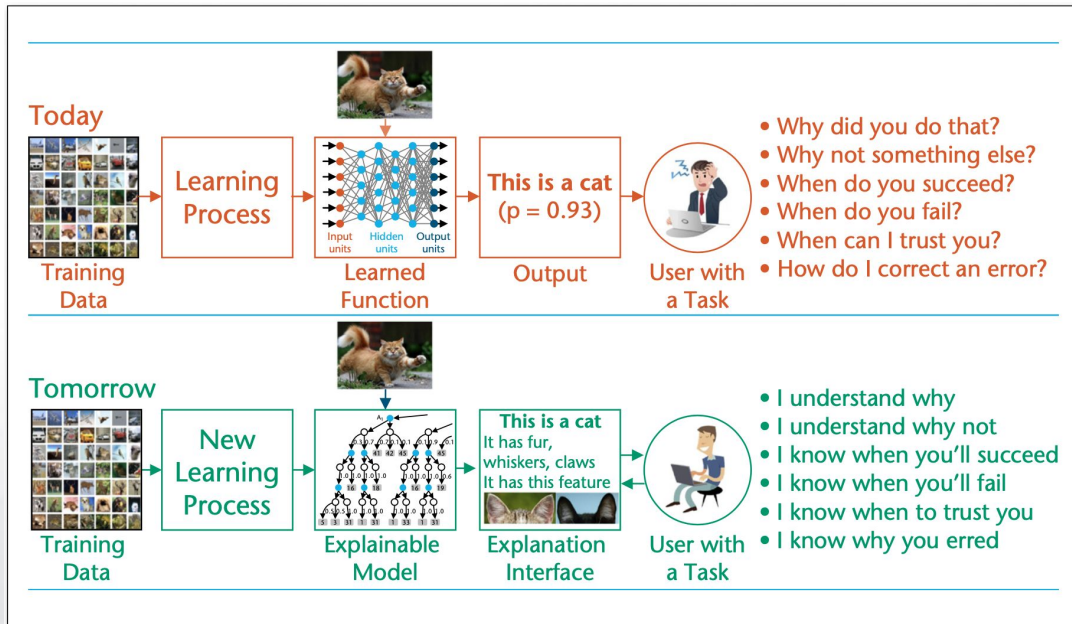
explanation | ɛksplə'neɪʃ(ə)n |

Oxford Dictionary of
English

A statement, fact, or situation that tells you why something happened;
a reason given for something.

What is explainable AI?

Explainable Artificial Intelligence (XAI) provides insight into the **why** for model predictions, offering potential for users to better understand and trust a model, and to recognize and correct AI predictions that are incorrect.



- AI provides explanations supporting its predictions.
- Explanations must be understandable to non-specialists.

Source: DARPA's Explainable Artificial Intelligence Program David Gunning, David W. Aha

What is explainable AI?

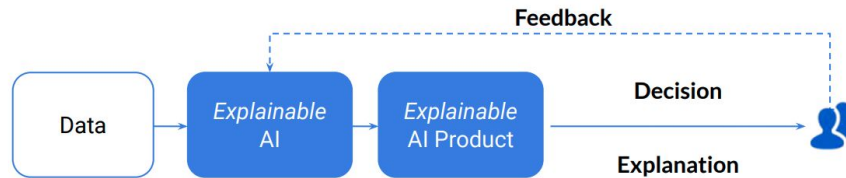
Black Box AI



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Explainable AI

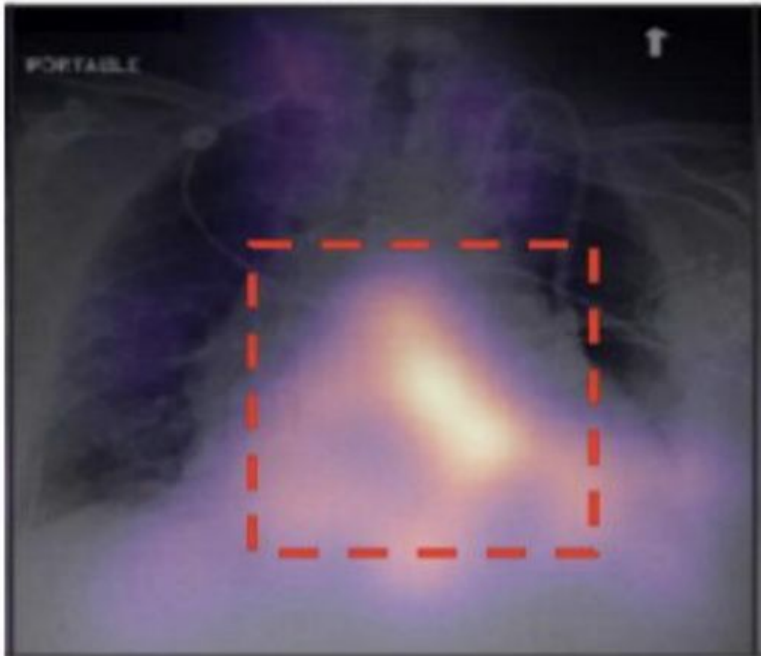


Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

A typical architecture of an explainable AI system

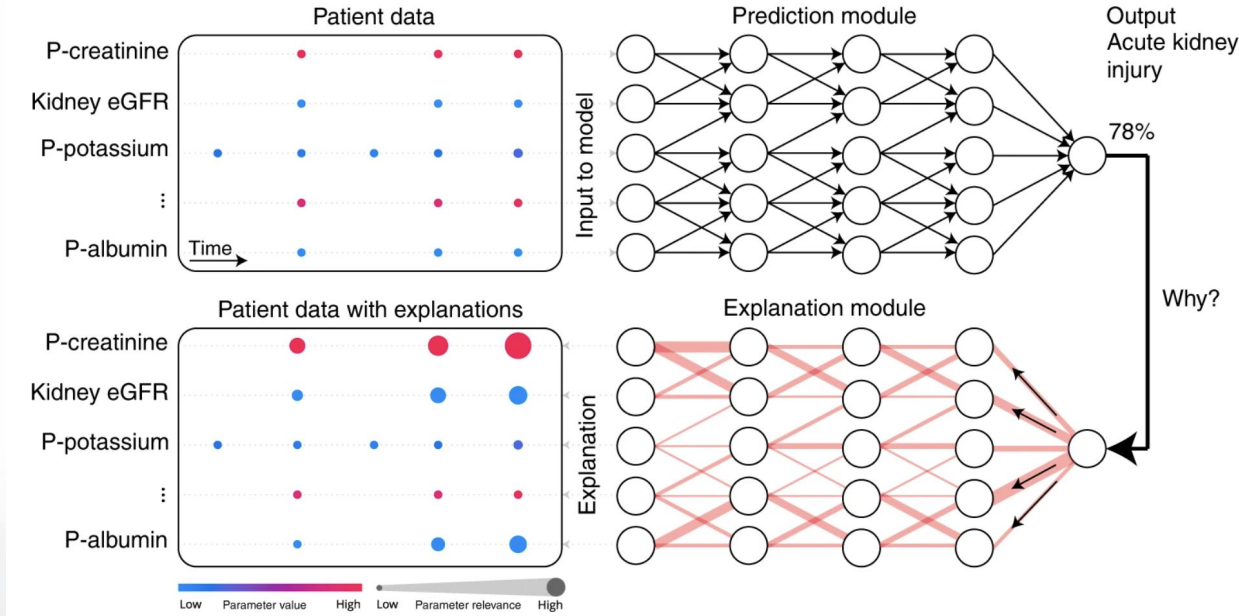
Explainable AI system: Example 1



VinDr AI platform predicts the presence of patient with cardiomegaly disease. Figure is taken from VinBigdata.

- An AI model predicts “cardiomegaly” with 95% confidence.
- Produce heatmaps [GradCAM, GBP, Class Activation Maps - CAM] for verifying the features learned by the proposed AI model.

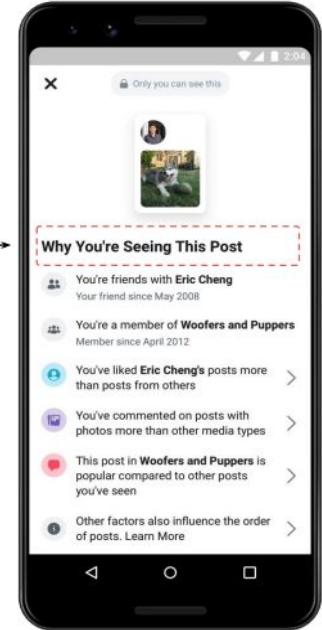
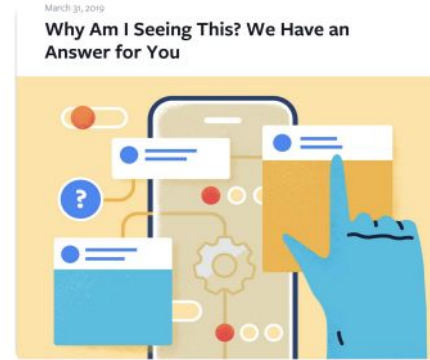
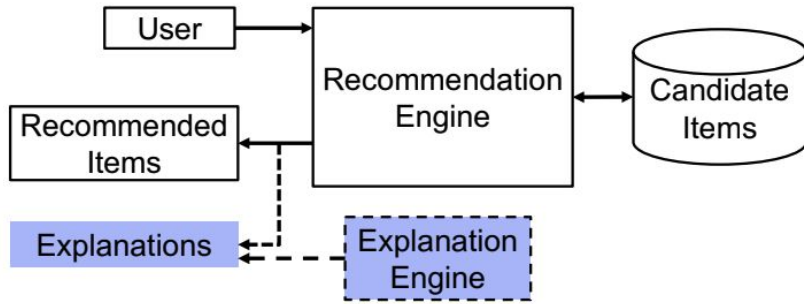
Explainable AI system: Example 2



- An explainable artificial intelligence early warning system for predicting acute critical illness from EHR.
- The explanation module then explains the TCN predictions in terms of input variables.

Lauritsen, Simon Meyer, et al. "Explainable artificial intelligence model to predict acute critical illness from electronic health records." *Nature Communications* 11.1 (2020): 1-11.


Explainable AI system: Example 3



Source: Facebook

Why Explainable AI is an Existential Need?

Black-box AI creates business risk for Industry: Wrong decision can be costly and dangerous.


 **Stanford**
MEDICINE | News Center

Menu

[Stanford Medicine / News](#) / Ethics review needed for AI use in health care


Researchers say use of artificial intelligence in medicine raises ethical questions


In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients

 Missouri S&T News and Research

After Uber, Tesla incidents, can artificial intelligence be trusted?


Apr 10, 2018




 **BBC NEWS**

Tay: Microsoft issues apology over racist chatbot fiasco


Sep 22, 2017



 **MIT News**

Study finds gender and skin-type bias in commercial AI systems

Feb 12, 2018

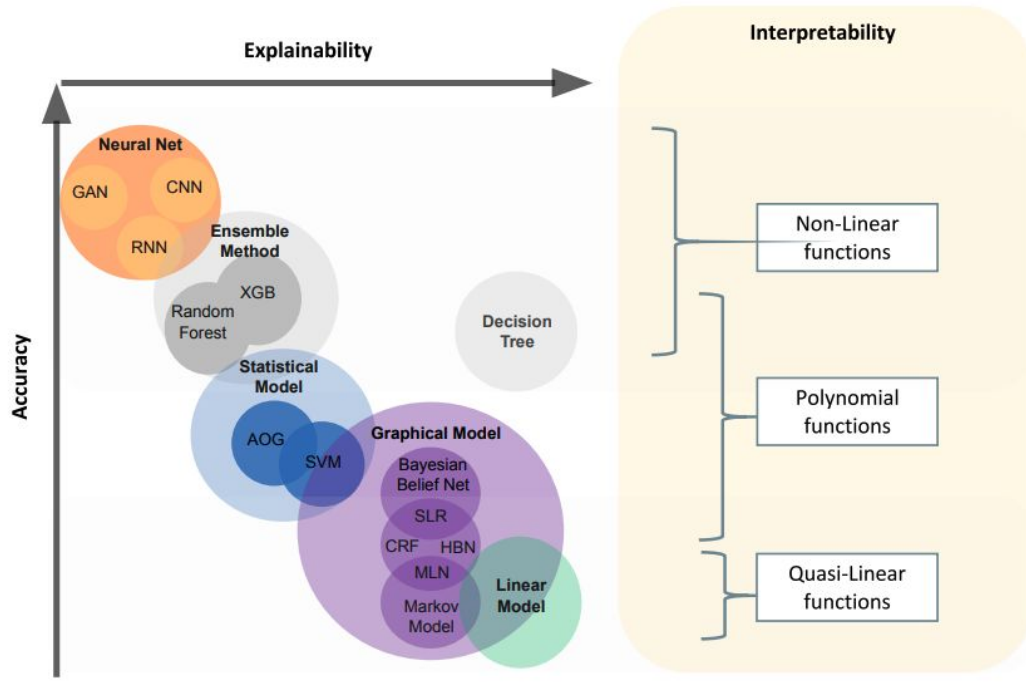


Why Explainable AI is an Existential Need?

Understanding a model's prediction is a critical for many tasks

- Explain predictions to support decision-making
- Recognize AI predictions that are correct/incorrect
- Verify that model behavior is acceptable
- And so on.

Accuracy vs Interpretability Trade-Offs



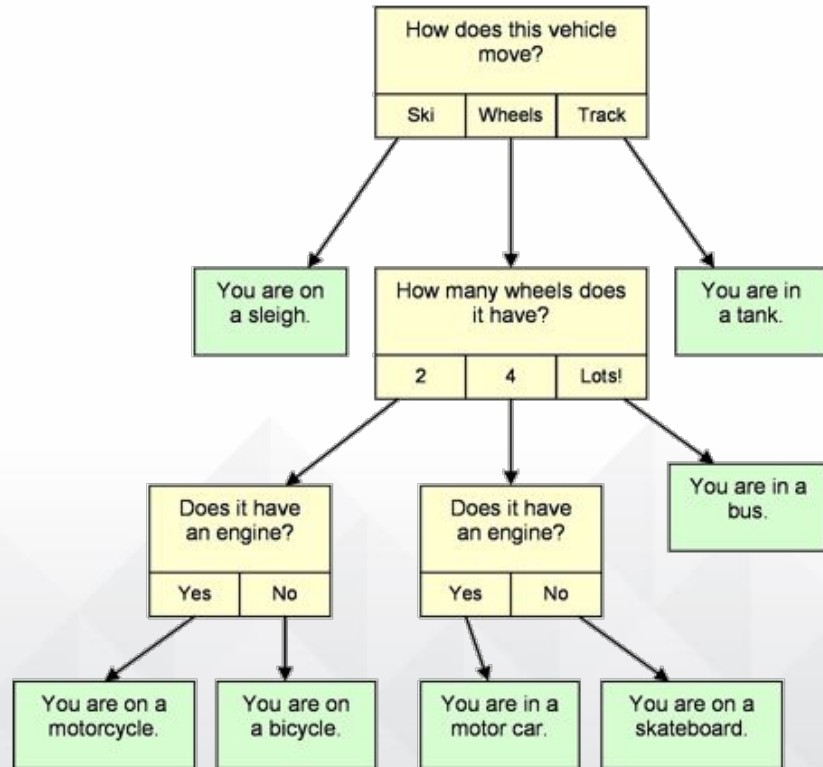
- The most accurate methods, such as convolutional neural nets (CNNs), provide no explanations;
- Understandable methods, such as rule-based, tend to be less accurate.

Freddy Lecue, et al. (AAAI 2020 Tutorial)

Accuracy and Interpretability Trade-Offs

Expert system:

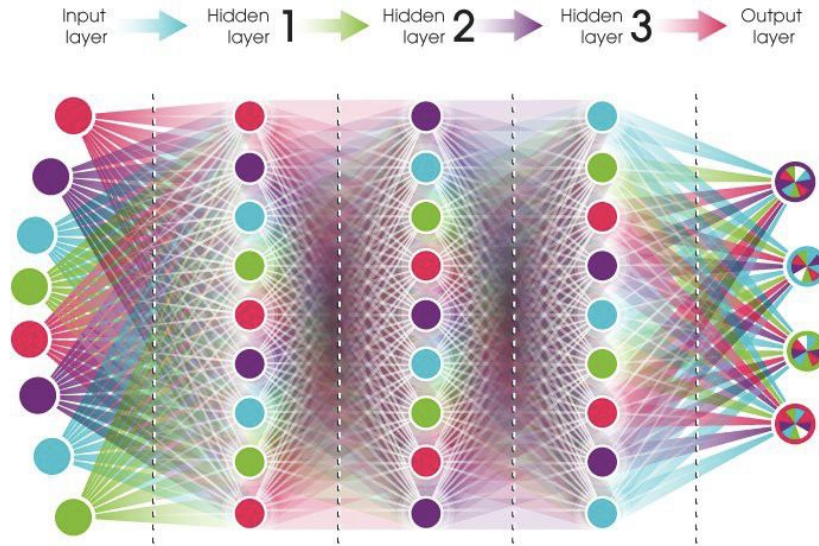
Good for explanations,
not so good for accuracy



The figure is taken from public domain.

Accuracy and Interpretability Trade-Offs

DEEP NEURAL NETWORK



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

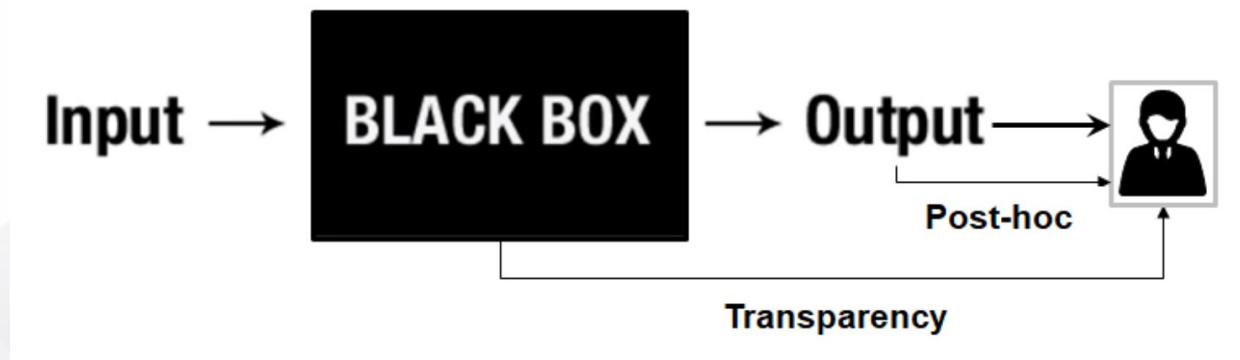
Deep Neural Networks:

Good for accuracy,
not so good for explanations

The figure is taken from public domain.

Interpretable machine learning methods

XAI techniques are classified in two categories of **transparent** and **post-hoc** methods.



Xu, Feiyu, et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges." CCF international conference on natural language processing and Chinese computing. Springer, Cham, 2019.

Interpretable machine learning methods

Transparent methods

- Transparent methods are such methods where the inner working and **decision-making process of the model is simple to interpret and represent.**
- These methods are useful where internal feature correlations are not that much complex or linear in nature.

E.g. Bayesian model, decision trees, linear regression, and fuzzy inference systems are examples of transparent models.

Interpretable machine learning methods

Transparent methods: An example

- Interpretable models – e.g. rule-based expert systems: “if patient has **symptoms A and B**, or **has B with C and D**, then illness is **X**”
 - best for explanations
 - hard to find optimal rules
 - less accurate than other approaches

Interpretable machine learning methods

Post-hoc methods

- Explain what the model has learned when it is not following a simple relationship among data and features.
- A post-hoc XAI method receives a trained and/or tested AI model as input, then generates useful approximations of the model's inner working and decision logic by producing **understandable representations in the form of feature importance scores, rule sets, heat maps, etc.**

Interpretable machine learning methods

Post-hoc methods: Examples

Showing the relationship between prediction and learned features.



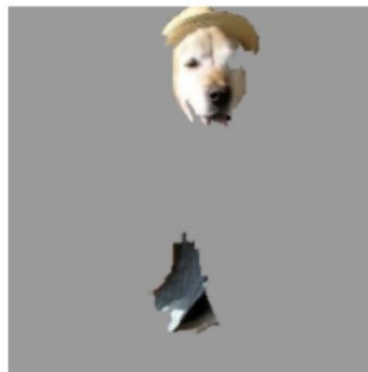
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Xu, Feiyu, et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges." CCF international conference on natural language processing and Chinese computing. Springer, Cham, 2019.

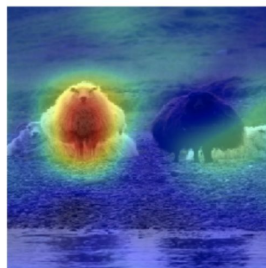
Interpretable machine learning methods

Post-hoc methods: Examples

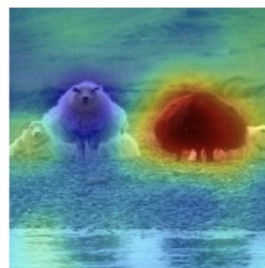
Showing the relationship between prediction and learned features.



(a) Sheep - 26%, Cow - 17%



(b) Importance map of 'sheep'



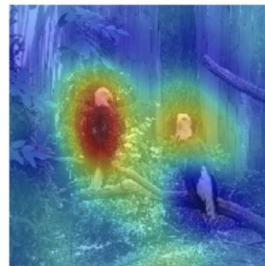
(c) Importance map of 'cow'



(d) Bird - 100%, Person - 39%



(e) Importance map of 'bird'



(f) Importance map of 'person'

Source:

<https://www.pcmag.com/news/the-next-step-toward-improving-ai>

Interpretable machine learning methods

Challenges



Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes (Chaofen Chen, Duke University).

Interpretable machine learning methods

Challenges

Do Vision Transformers See Like Convolutional Neural Networks?

Maithra Raghu¹ Thomas Unterthiner¹ Simon Kornblith¹ Chiyuan Zhang¹ Alexey Dosovitskiy¹

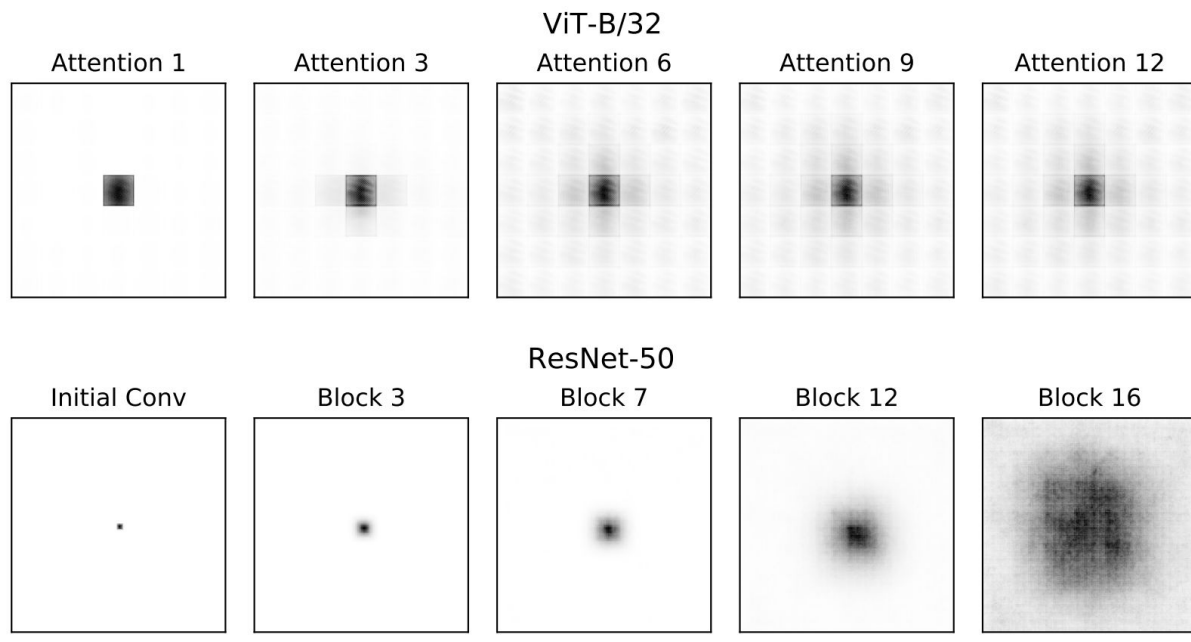
¹*Google Research, Brain Team*

Abstract

Convolutional neural networks (CNNs) have so far been the de-facto model for visual data. Recent work has shown that (Vision) Transformer models (ViT) can achieve comparable or even superior performance on image classification tasks. This raises a central question: *how are Vision Transformers solving these tasks?* Are they acting like convolutional networks, or learning entirely different visual representations? Analyzing the internal representation structure of ViTs and CNNs on image classification benchmarks, we find striking differences between the two architectures, such as ViT having more uniform representations across all layers. We explore how these differences arise, finding crucial roles played by self-attention, which enables early aggregation of global information, and ViT residual connections, which strongly propagate features from lower to higher layers. We study the ramifications for spatial localization, demonstrating ViTs successfully preserve input spatial information, with noticeable effects from different classification methods. Finally, we study the effect of (pretraining) dataset scale on intermediate features and transfer learning, and conclude with a discussion on connections to new architectures such as the MLP-Mixer.

Interpretable machine learning methods

Challenges



- There are striking differences between the two models.
- Vision Transformer has more uniform representations, with greater similarity between lower and higher layers.

Summary

- Explainability is a critical problem in the acceptance of artificial intelligence/machine learning, especially for critical applications.
- Human users will not trust AI if conclusions cannot be explained.
- Need to have more and better explainable AI models.

References

Vodrahalli, Kailas, Tobias Gerstenberg, and James Zou. "Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions." *arXiv preprint arXiv:2107.07015* (2021).

Doran, Derek, Sarah Schulz, and Tarek R. Besold. "What does explainable AI really mean? A new conceptualization of perspectives." *arXiv preprint arXiv:1710.00794* (2017).

Hoffman, Robert R., et al. "Metrics for explainable AI: Challenges and prospects." *arXiv preprint arXiv:1812.04608* (2018).

Explainable AI: The next stage of human-machine collaboration. URL:
<https://www.accenture.com/us-en/insights/technology/explainable-ai-human-machine>

Holzinger, Andreas. "From machine learning to explainable AI." 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE, 2018.

Holzinger, Andreas, et al. "What do we need to build explainable AI systems for the medical domain?." *arXiv preprint arXiv:1712.09923* (2017).



**VinUni-Illinois Smart Health Center
Call for Applications for Research Assistants**

I am looking for self-motivated Ph.D. and Research Assistants, who are interested in artificial intelligence for smart healthcare, computer vision, machine/deep learning, and natural language processing. If you are interested in doing research with me, please send me an email (hieu.ph@vinuni.edu.vn) with your CV and transcripts.

Thank you!