

Advanced Topics in Artificial Intelligence & Machine Learning

Interpretability, Explainability, and AI Ethics

Aleksandr Petiushko

ML Research

March 11th, 2024



Content

① Interpretability

Content

- 1 Interpretability
- 2 Explainability

Content

- 1 Interpretability
- 2 Explainability
- 3 Bias and Fairness in AI

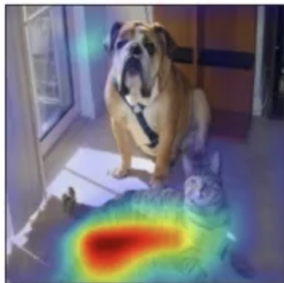
Content

- 1 Interpretability
- 2 Explainability
- 3 Bias and Fairness in AI
- 4 AI Ethics

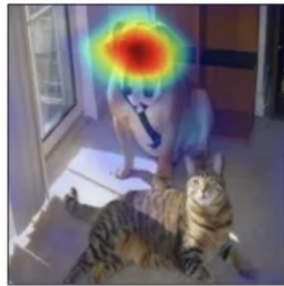
Interpretability

- Interpretability: understand the influence of any input sub-area/sub-feature on the model output;
- Can be understood as a sophisticated tool towards the ML Debug system
- Can be done via input counterfactual analysis (changing/reverting some input features)
- Read material: [link](#)

Grad-CAM for "Cat"

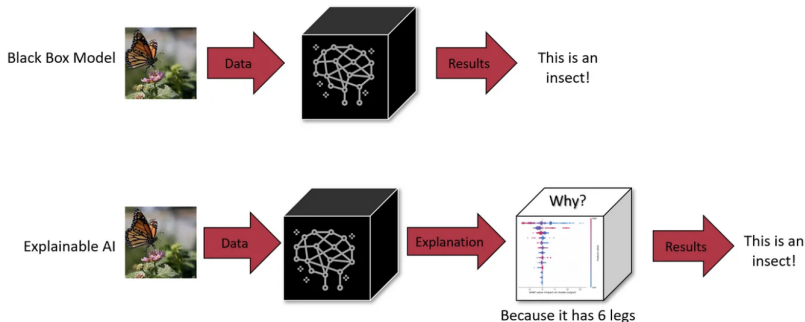


Grad-CAM for "Dog"



Explainability

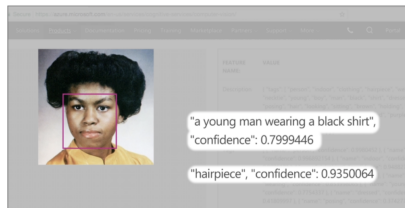
- Explainability: high-level interpretability (using human-like language), a clear and intuitive explanation of the decisions made
- Explainability now can be done via LLM chain-of-thought (CoT) technique
- Read material: [link](#)



Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)
- ML model can provide biased predictions w.r.t. race, gender, age. etc
- The main reason: skewed/sparse training data
- Main technique to avoid: human-in-the-loop, alignment
- Read material: [link](#)

Michelle Obama



AI ethics and regulations¹

Inequity and fairness

ML can contribute to and amplify social **inequity**

For **foundation models**, it is useful to separate:

- **intrinsic biases** (properties in the foundation model)
- **extrinsic harms** (harms in specific applications)

Source tracing to understand ethical/legal responsibility

Mitigations: **proactive interventions**/**reactive recourse**

Environment

Foundation models involve significant training/**emissions**

One perspective: **amortised** cost over re-use

Several factors would be **beneficial** to consider:

- **compute-efficient models**, **hardware**, **energy grids**
- **environmental cost** as a factor for evaluation
- greater **documentation** and measurement

Economics

Foundation models may have **economic impact** due to:

- **novel capabilities**
- potential applications in **wide array of industries**

Initial analyses have been conducted to understand implications for **productivity**, **wage inequality**, **concentration of ownership**

Misuse

Misuse: the use of foundation models as technically intended but for societal harm (e.g. disinformation)

Foundation models may make misuse easier by generating **high-quality** personalised content

Disinformation actors can target demographic groups

Foundation models may also help to **detect misuse**

Legality

How **law** bears on development/deployment is unclear

Legal/regulatory frameworks will be needed

In the **US** setting, important issues include:

- **liability** for model predictions
- **protections** from model behaviour

Legal standards must advance for intermediate models

Ethics of scale

Widespread adoption of foundation models poses ethical, political and social concerns

Ethical issues related to **scale**:

- **homogenisation**
- **concentration of power**

How can **norms** and **release strategies** address these?

¹www.youtube.com

Takeaway notes

- 1 Read all the mentioned links

Takeaway notes

- ① Read all the mentioned links
- ② Interpretability and Explainability are quite connected in ML

Takeaway notes

- ① Read all the mentioned links
- ② Interpretability and Explainability are quite connected in ML
- ③ Interpretability deals mostly on a lower level, input/output dependencies

Takeaway notes

- ① Read all the mentioned links
- ② Interpretability and Explainability are quite connected in ML
- ③ Interpretability deals mostly on a lower level, input/output dependencies
- ④ Explainability steps in on a higher level to provide a human-like explanations

Takeaway notes

- ➊ Read all the mentioned links
- ➋ Interpretability and Explainability are quite connected in ML
- ➌ Interpretability deals mostly on a lower level, input/output dependencies
- ➍ Explainability steps in on a higher level to provide a human-like explanations
- ➎ Usually the most interpretable are simpler models; explainability can be applied to a model of any complexity

Thank you!