Explaining Neural Decision-making: Model-understanding Tools and Interpretability Techniques

Alexander J. Hepburn

@ *a.hepburn.1@research.gla.ac.uk*

@_ajhepburn

University of Glasgow

What is interpretability and what makes it valuable for research?

> “Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human.” (Doshi-Velez and Kim, 2017)

Interpretability is important because...

- It builds the groundwork for model debugging
- It can inform decision-making
- It can facilitate improved data collection and feature engineering techniques
- It has a wide range of applicability in machine learning research

What is interpretability and what makes it valuable for research?

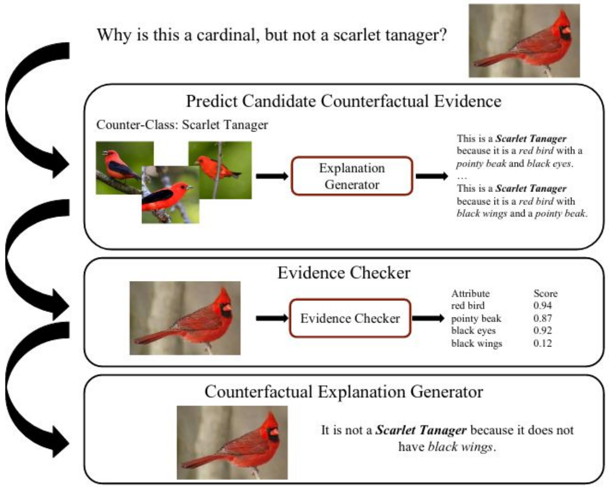
> “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim, 2017)

It allows us to answer...

- What does my model perform poorly on?
- Why did my model make this prediction?
- Does my model behave consistently if I modify the linguistic constraints of my document? (ie. grammar, tense, pronouns)

What techniques are used in the interpretation of language models?

- Feature-, Neuron-, Layer-importance algorithms
- Counterfactual generation
- Attention visualisation



Why is this a cardinal, but not a scarlet tanager?

Predict Candidate Counterfactual Evidence

Counter-Class: Scarlet Tanager

Explanation Generator

This is a ***Scarlet Tanager*** because it is a *red bird* with a *pointy beak* and *black eyes*.

…

This is a ***Scarlet Tanager*** because it is a *red bird* with *black wings* and a *pointy beak*.

Evidence Checker

Evidence Checker

Attribute	Score
red bird	0.94
pointy beak	0.87
black eyes	0.92
black wings	0.12

Counterfactual Explanation Generator

It is not a ***Scarlet Tanager*** because it does not have *black wings*.

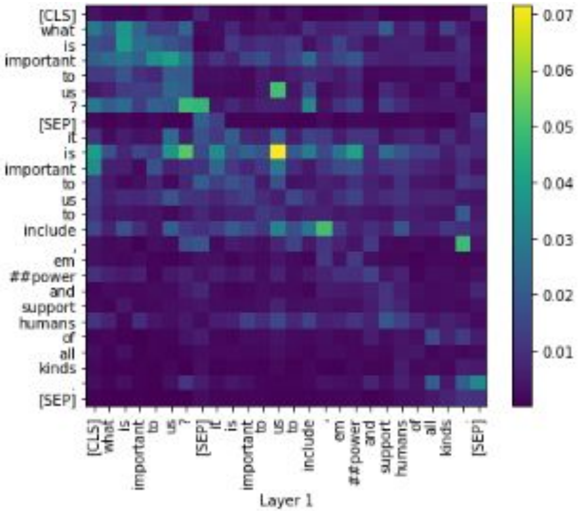
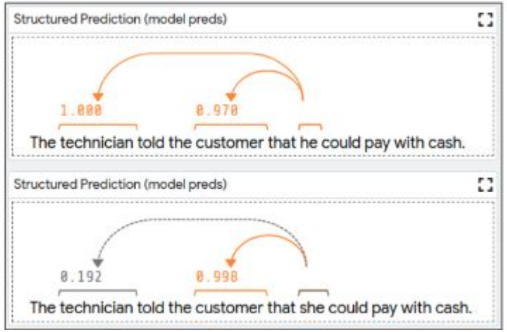


Fig. 1: Example of counterfactual explanation. *Generating Counterfactual Explanations with Natural Language*. Hendricks et al. (2018).
Fig. 2: Visualising self-attention in BERT. *Captum - Interpreting BERT Models (Part 2)*. https://captum.ai/tutorials/Bert_SQUAD_Interpret. Accessed: 14th April 2021.

What are some common applications of interpretability?

- Image classification (Kokhlikyan et al., 2019)
- Debugging text generation (Strobert et al., 2018; Tenney et al., 2020)
- Coreference resolution (Tenney et al., 2020)
- Explainable recommendation (Pan et al., 2020)


Structured Prediction (model preds)
1.000
0.970

The technician told the customer that he could pay with cash.

Structured Prediction (model preds)
0.192
0.998

The technician told the customer that she could pay with cash.

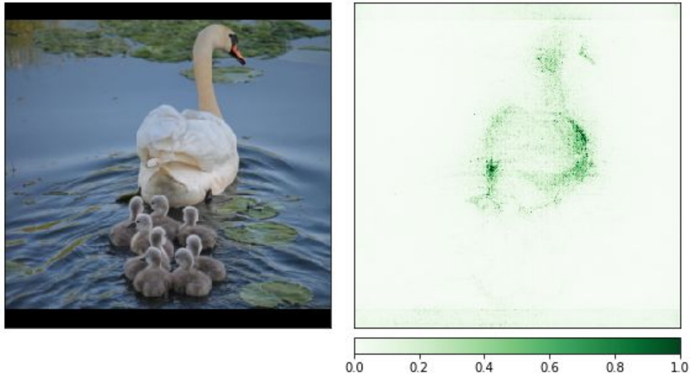
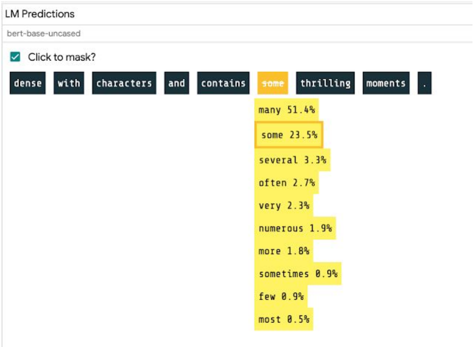


Fig. 1: Example of LIT’s coreference resolution visualisation. *The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models.* Tenney et al. (2020).

What tools exist that employ these techniques?

- InterpretML
- AllenNLP Interpret
- exBERT
- SHAP
- **LIME**
- **Language Interpretability Tool (LIT)**
- **Captum**

Captum



LM Predictions

bert-base-uncased

☑ Click to mask?

dense with characters and contains ~~some~~ thrilling moments .

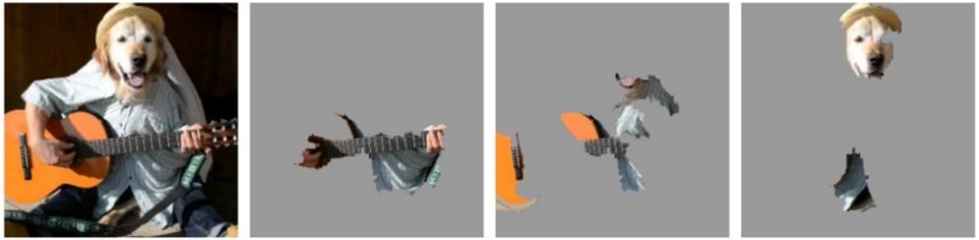
```
many 51.4%
some 23.5%
several 3.3%
often 2.7%
very 2.3%
numerous 1.9%
more 1.8%
sometimes 0.9%
few 0.9%
most 0.5%
```

LIME

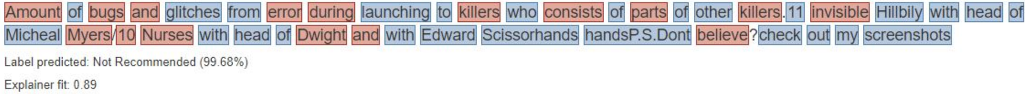
- Used to explain a variety of different black-box classification processes via **perturbations** of the interpretable instance

- Model agnostic

- Provides **local** explanations of a surrounding input sequence



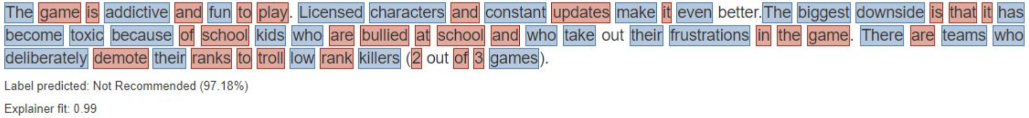
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*



Amount of bugs and glitches from error during launching to killers who consists of parts of other killers.11 invisible Hillbily with head of Micheal Myers/10 Nurses with head of Dwight and with Edward Scissorhands handsP.S.Dont believe?check out my screenshots

Label predicted: Not Recommended (99.68%)

Explainer fit: 0.89



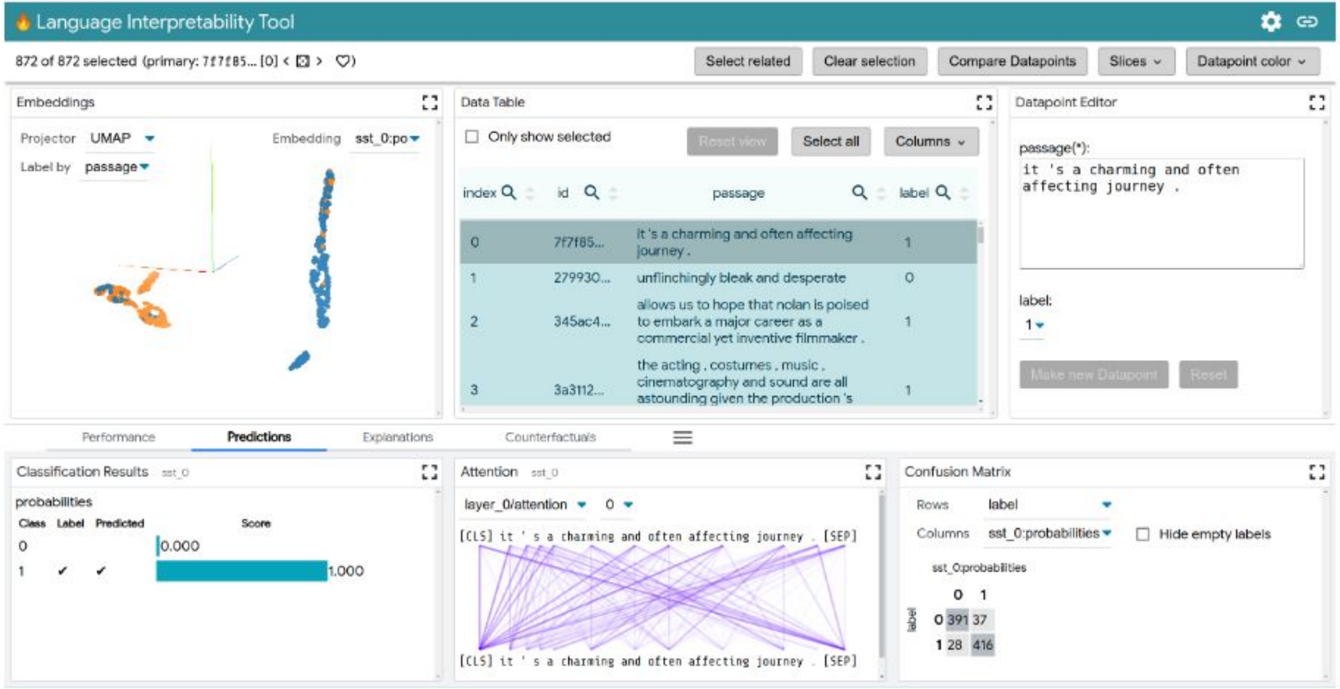
The game is addictive and fun to play. Licensed characters and constant updates make it even better.The biggest downside is that it has become toxic because of school kids who are bullied at school and who take out their frustrations in the game. There are teams who deliberately demote their ranks to troll low rank killers (2 out of 3 games).

Label predicted: Not Recommended (97.18%)

Explainer fit: 0.99

Fig. 1, 2: Example of image and text classification explanations. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Ribeiro et al. (2016).

Language Interpretability Tool (LIT)



872 of 872 selected (primary: 7f7f85... [0] < ⊡ > ♡)

□ Only show selected

index	id	passage	label
0	7f7f85...	it 's a charming and often affecting journey .	1
1	279930...	unflinchingly bleak and desperate	0
2	345ac4...	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	3a3112...	the acting , costumes , music , cinematography and sound are all astounding given the production 's	1

passage(*):

```
it 's a charming and often
affecting journey .
```

label:
1

probabilities

Class	Label	Predicted	Score
0			0.000
1	✓	✓	1.000

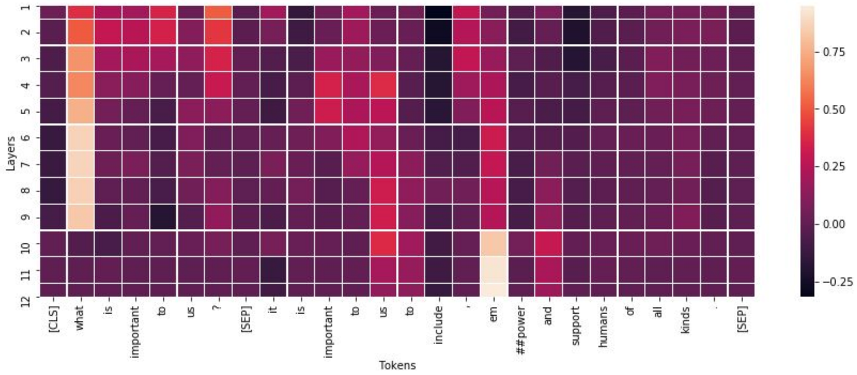
sst_0:probabilities

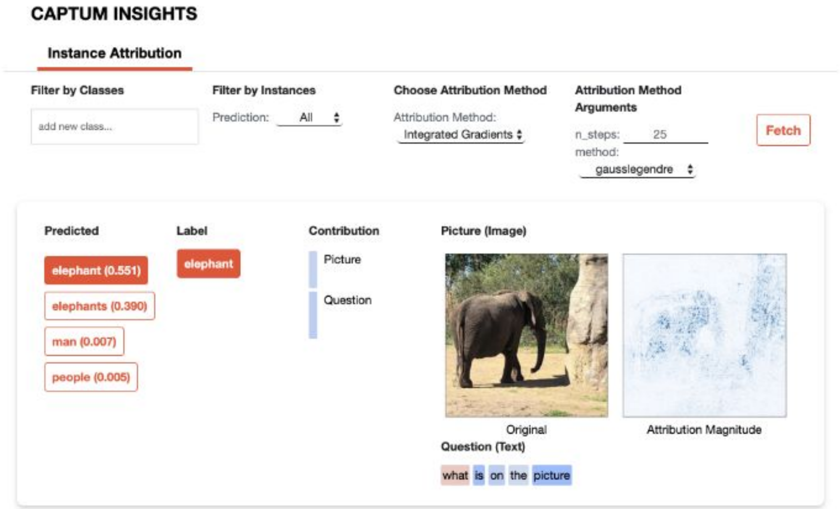
label	0	1
0	391	37
1	28	416

Fig. 1: Example of LIT's UI. *The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models.* Tenney et al. (2020).

Captum

- Interpretability library for PyTorch (developed by FAIR).
- Provides a more **complete** collection of interpretability techniques
- Multi-GPU support
- Not as straightforward to use!




CAPTUM INSIGHTS

Instance Attribution

Filter by Classes: add new class...

Filter by Instances — Prediction: All

Choose Attribution Method — Attribution Method: Integrated Gradients

Attribution Method Arguments — n_steps: 25 — method: gausslegendre

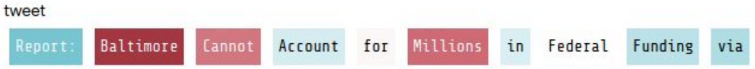
Question (Text): what is on the picture

Fig. 1, 2: Example of layer conductance, image attribution UI. *Captum: A unified and generic model interpretability library for PyTorch*. Kokhlikyan et al. (2020).

Debugging crisis- and disaster-related social media content

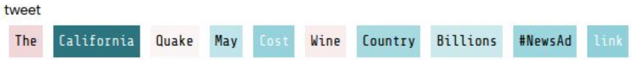
- Able to provide local explanations of **task-specific** attributions
- Allowed us to **debug** where our model was failing on a local, per-token level
- Allowed us to identify **incorrectly labelled** samples in the TREC-IS dataset

Incorrectly Labelled: **Irrelevant**



tweet

Report: Baltimore Cannot Account for Millions in Federal Funding via

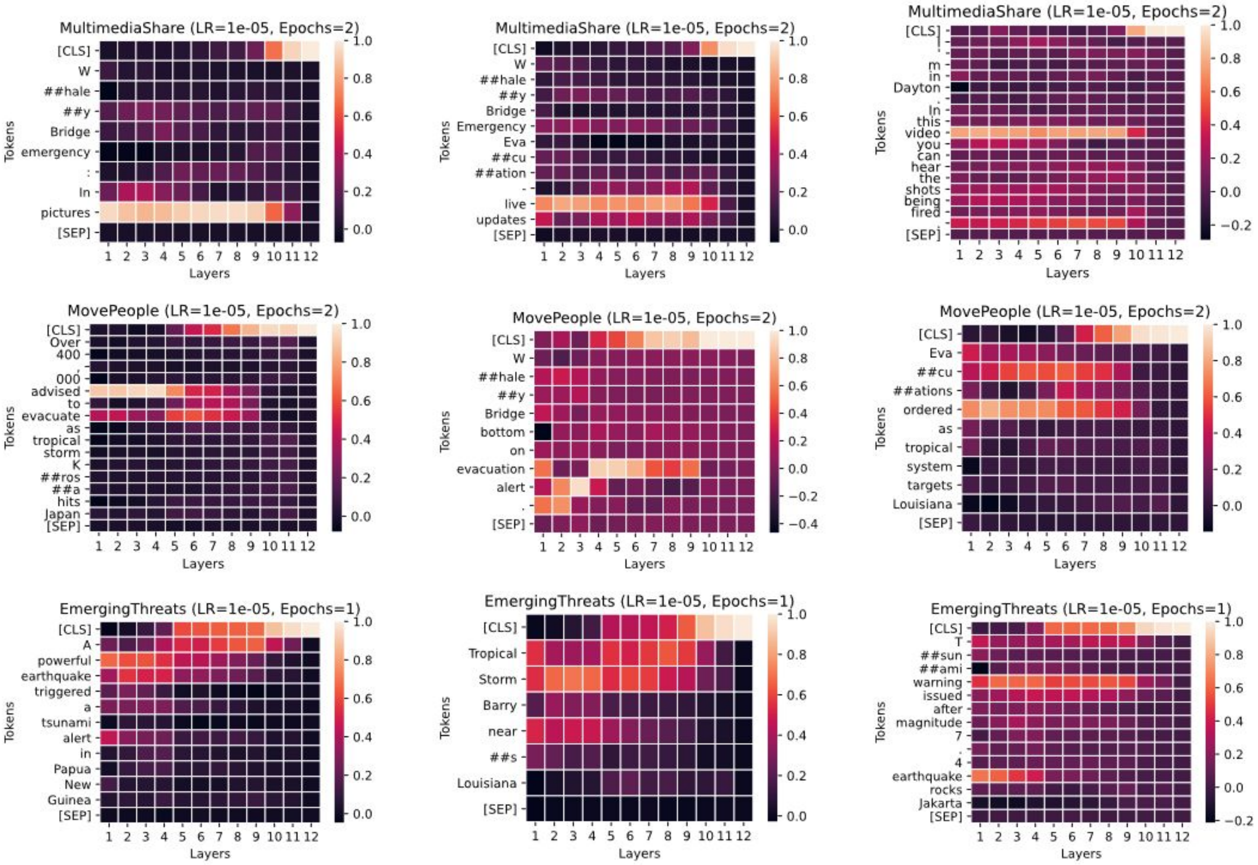


tweet

The California Quake May Cost Wine Country Billions #NewsAd link

Correctly Labelled: **Location**

What items do our classifiers pay particular attention to?



MultimediaShare (LR=1e-05, Epochs=2)

MultimediaShare (LR=1e-05, Epochs=2)

MultimediaShare (LR=1e-05, Epochs=2)

MovePeople (LR=1e-05, Epochs=2)

MovePeople (LR=1e-05, Epochs=2)

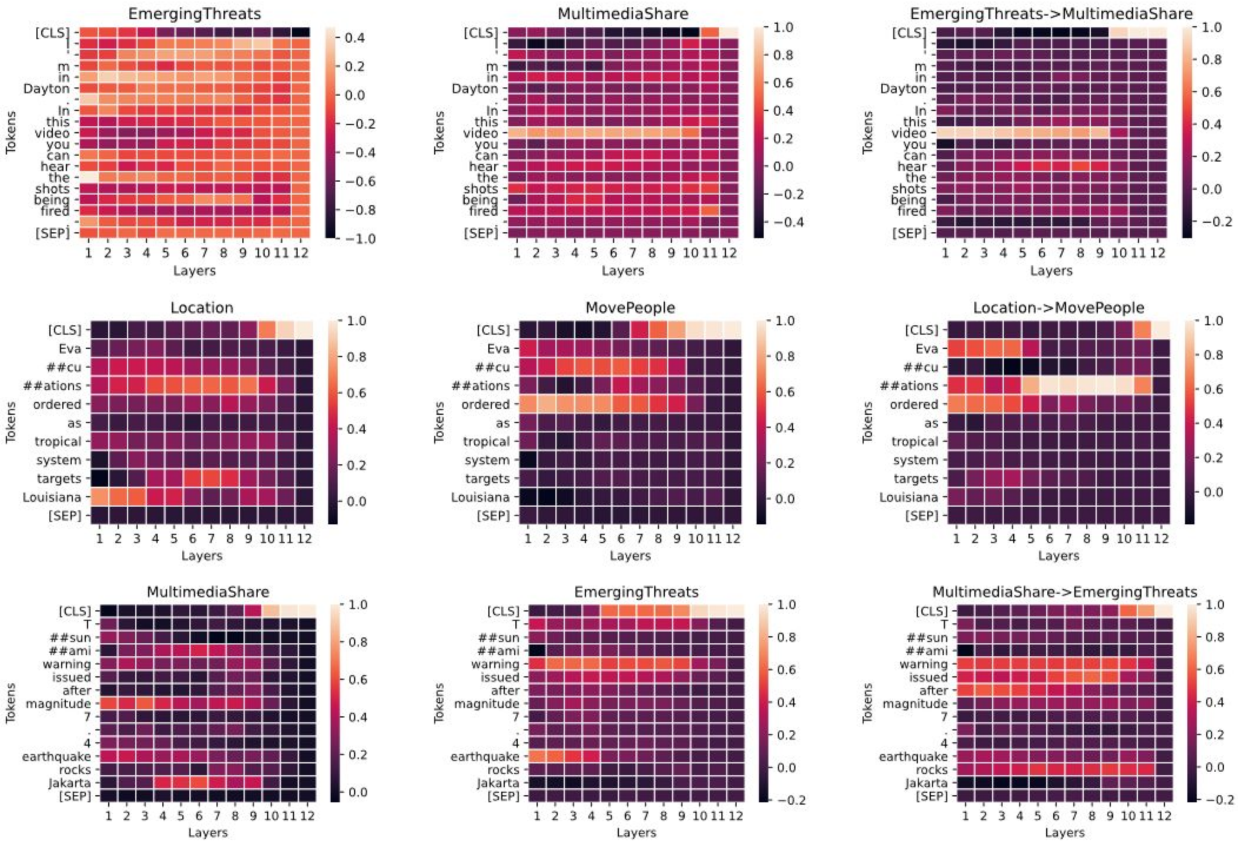
MovePeople (LR=1e-05, Epochs=2)

EmergingThreats (LR=1e-05, Epochs=1)

EmergingThreats (LR=1e-05, Epochs=1)

EmergingThreats (LR=1e-05, Epochs=1)

What changes when we train across related tasks in succession?



What are some recent, exciting developments in interpretability research?

- NNs are vulnerable to adversarial attacks (Ren et al., 2020), building more *trustworthy* systems is more important than ever.
- Theoretical groundwork of model explanation is forming (Arrieta et al., 2019) as field develops.
- Allows for model debugging. (Hall et al., 2019)

There are some known limitations...

- Current GUI-based tools do not scale well to large datasets. (ie. LIT, ~10k examples, Tenney et al., 2020)
- Gradient-based attributions are manipulable. (Kindermans et al., 2017; Wang et al., 2020)
- Attention mechanisms may not provide meaningful explanations (Jain and Wallace, 2019)

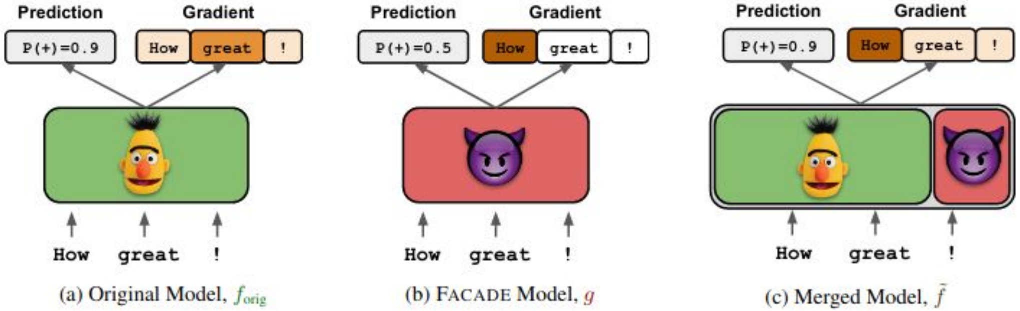


Fig. 1: Example of gradient-based attribution manipulation via the FACADE model. *Gradient-based Analysis of NLP Models is Manipulable*. Wang et al. (2020).

Thanks for listening!

Any questions?