Model Interpretation with Machine Learning and Interactive Visualization

Xiting Wang

Why Model Interpretation?

- Human intelligence (HI) and artificial intelligence (AI) should work together to achieve the best result
  - AI can really help get interesting insights for human and help decision making
  - Human knowledge learned based on years of experience and careful analysis of the data is quite valuable
- Explainable AI builds the bridge

Which Features Contribute to Prediction?

[SOS]	rare	bird	has	more	than	enough	charm	to	make	it	memorable	.

Measuring the input contribution is useful for

Understanding

Correct for the right reason?

Debug

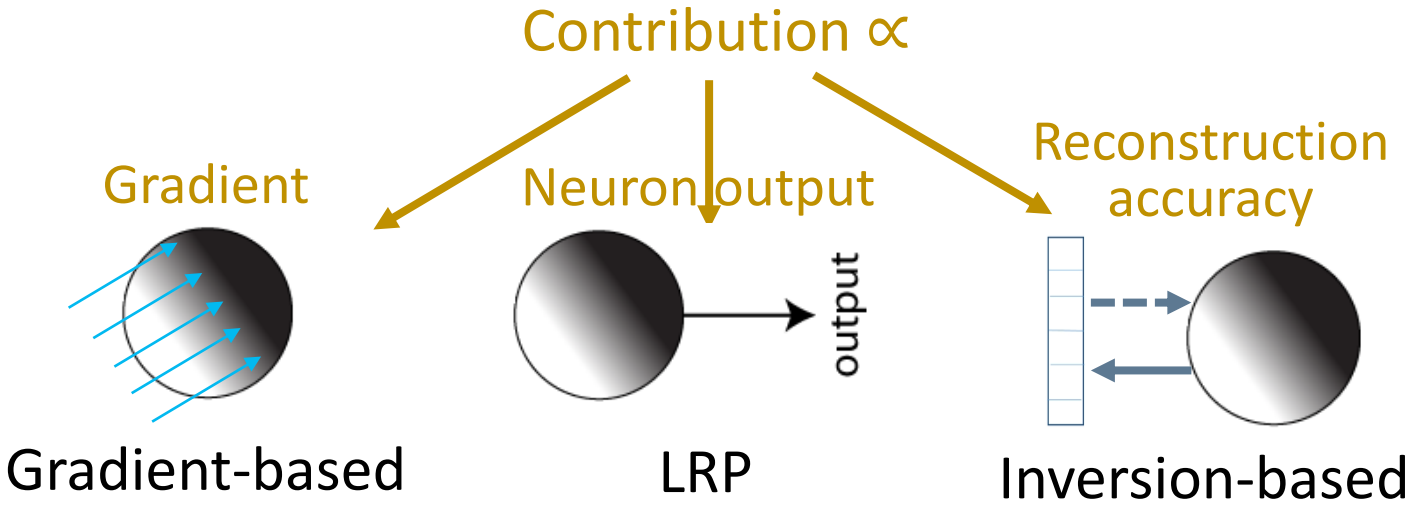
irrelevant features used?

Trust

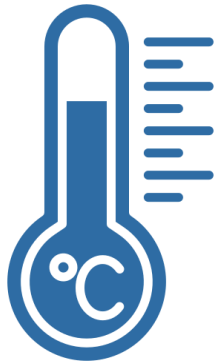
Safe and fair?

Challenge: Lack of A Unified Measure

- Existing works define contribution from one ***heuristic*** angle



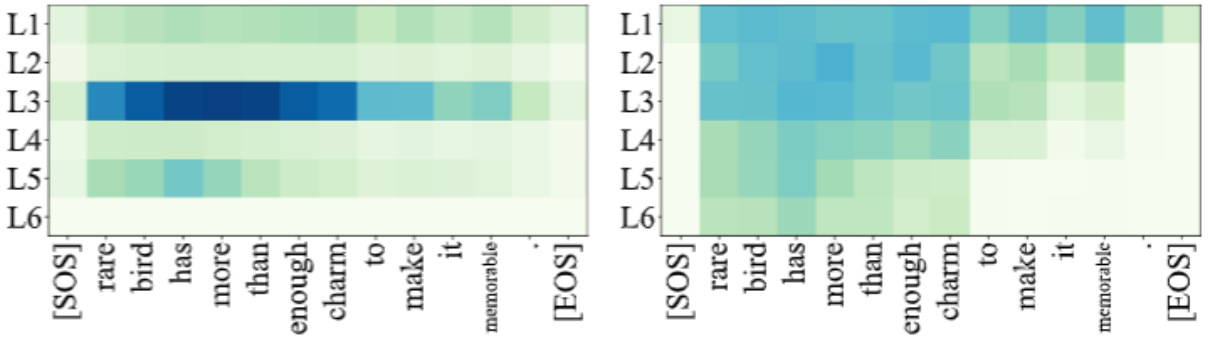
- Lack of a ***unified*** measure
  - ***Universality***
  - ***Coherency***



Our Unified Measure

We find a unified measure for model interpretation: *mutual Information (MI)*

- ***Universality:*** As a fundamental quantity in information theory, MI is can be defined for any model architectures and tasks
- ***Coherency:*** (Kinney et al. 2014) prove that MI quantifies associations without bias with respect to relationships of a specific form, enabling us to achieve ***consistent comparison*** across neurons, layers, and models



(a) Gradient-based method (b) Ours

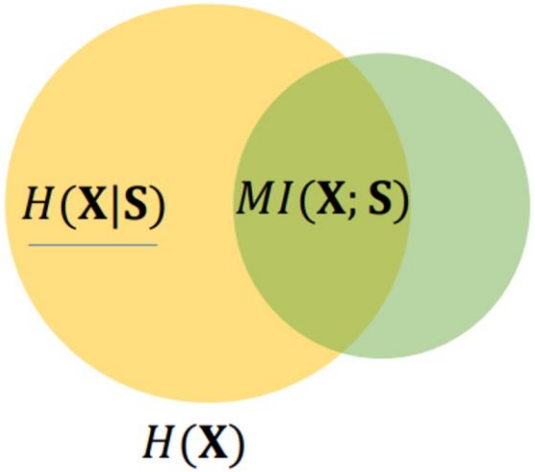
Our Unified Measure

We find a unified measure for model interpretation: *mutual Information (MI)*

Methods	Coherency			Universality
	Neuron	Layer	Model	
Gradient-based	✓	×	×	×
Inversion-based	✓	×	×	×
LRP	×	×	×	×
Ours	✓	✓	✓	✓

Multi-level Mutual Information Quantification

The contribution of input X
on hidden state S



$$\underbrace{MI(\boldsymbol{X};\boldsymbol{S})}_{\textbf{Mutual information}} = \underbrace{H(\boldsymbol{X})}_{\text{Constant}} - \underbrace{H(\boldsymbol{X}|\boldsymbol{S})}_{\textbf{Discarded information}}$$

Decomposing into the feature (word) level

Corpus level → **Sample** level → **Word** level

$\boldsymbol{H(X|S)}$ $\quad$ $\boldsymbol{H(X|s)}$ $\quad$ $\boldsymbol{H(X_i|s)}$

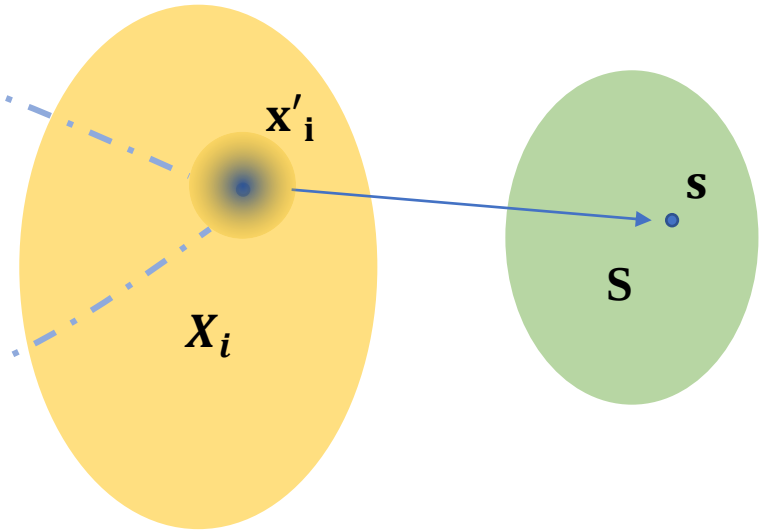
Perturbation-based Approximation

Compute the discarded word information $H(\mathbf{X}_i|\mathbf{s})$

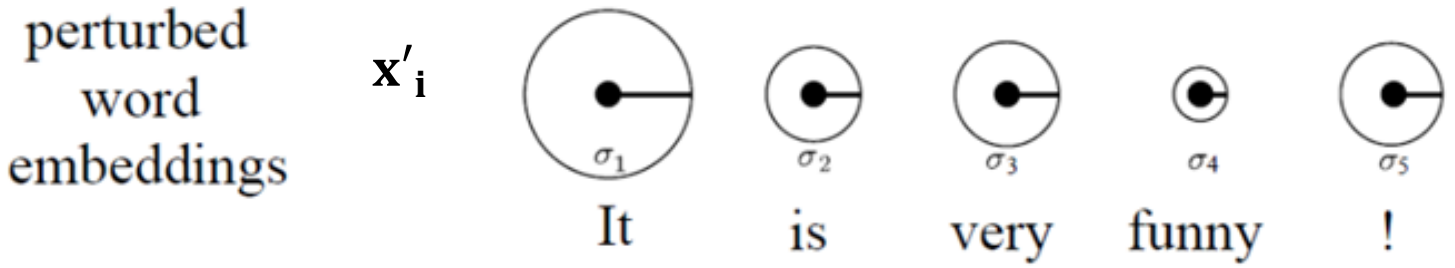
$$H(\mathbf{X}_i|\mathbf{s}) = -\int_{\mathbf{x}_i' \in \mathbf{X}_i} p(\mathbf{x}_i'|\mathbf{s}) \log p(\mathbf{x}_i'|\mathbf{s}) \, d\mathbf{x}_i'$$

Intractable ➡ Approximate with Gaussian

Suppose $\mathbf{x'_i} = \mathbf{x}_i + \boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{I})$



Learn σ_i^2 by 1) minimizing change in s with MLE;
2) maximizing entropy



Coherency Across Neuros

Task: reverse a sequence

a b c d e => e d c b a

t=2 a b c d e => e d c b a

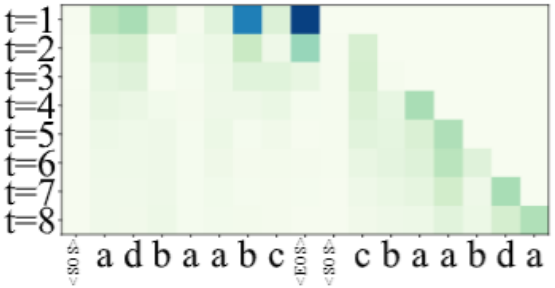
t=3 a b c d e => e d c b a

t=4 a b c d e => e d c b a

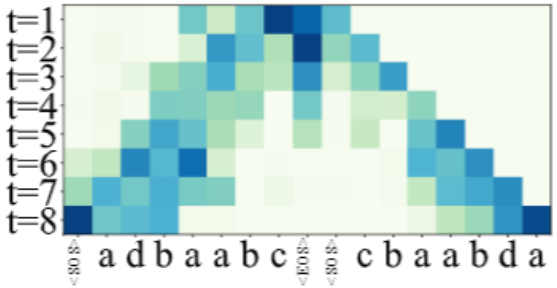
t=5 a b c d e => e d c b a

The reverse pattern

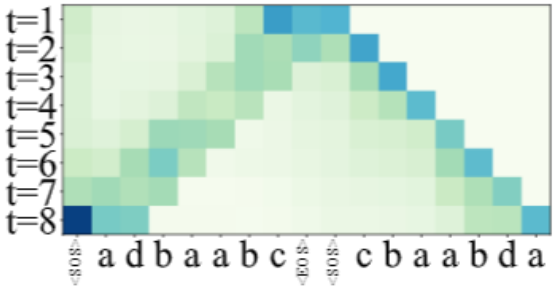
LRP



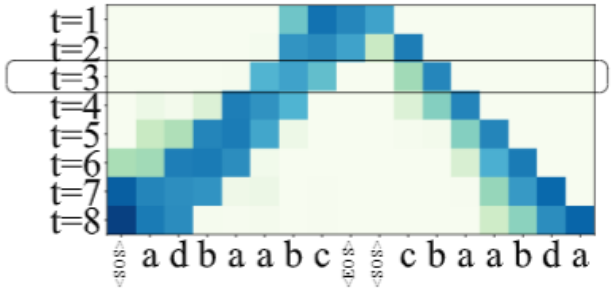
Perturbation



Gradient



Ours



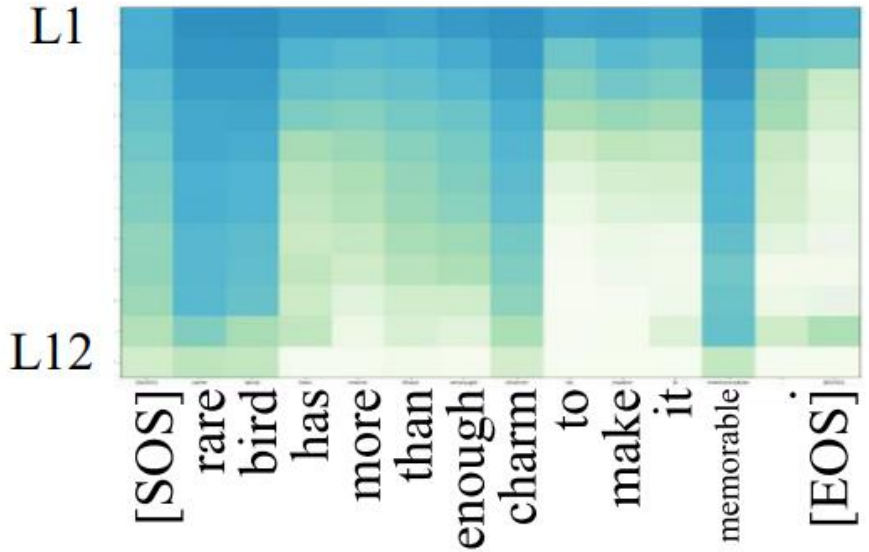
Our method shows the clearest reverse pattern

Coherency Across Layers

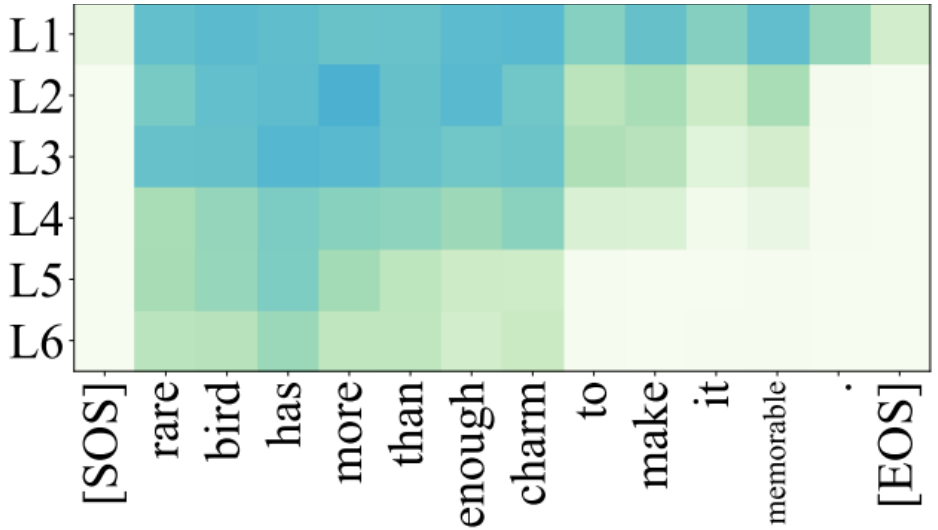
Our method clearly shows how the network gradually processes input words through layers

Task: sentiment classification

Interpret BERT with ours



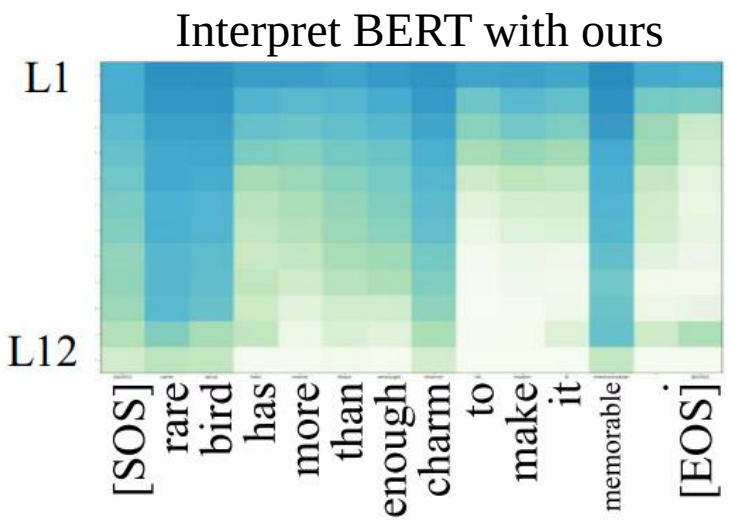
Interpret Transformer with ours

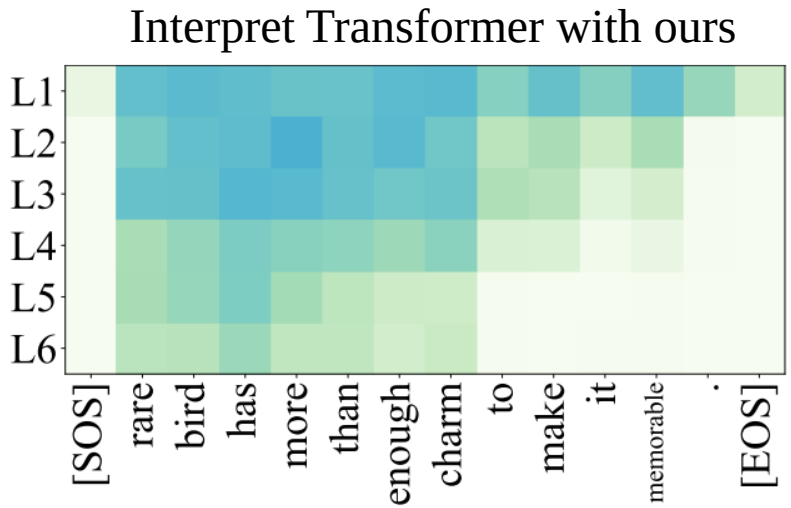


Coherency Across Layers

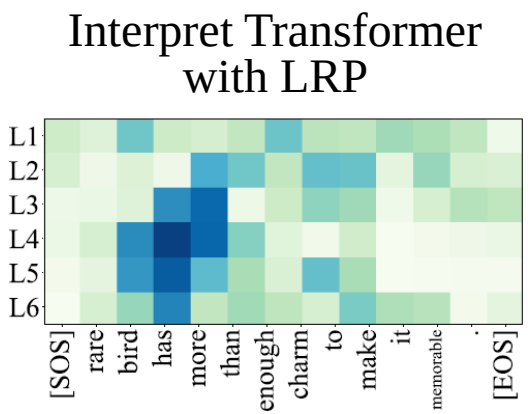
Our method clearly shows how the network gradually processes input words through layers

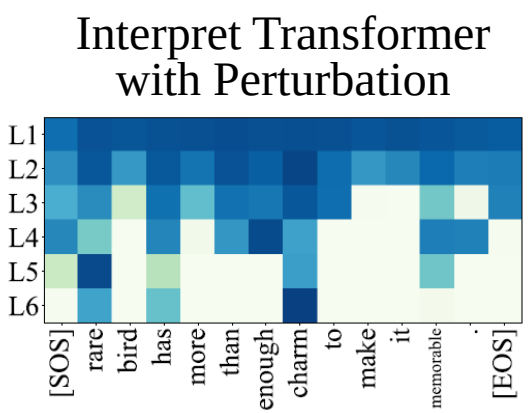
Ours

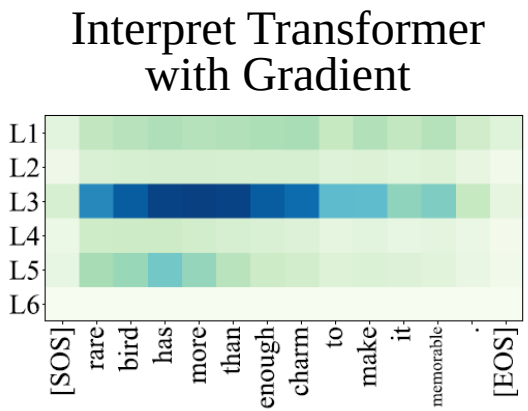




Baselines

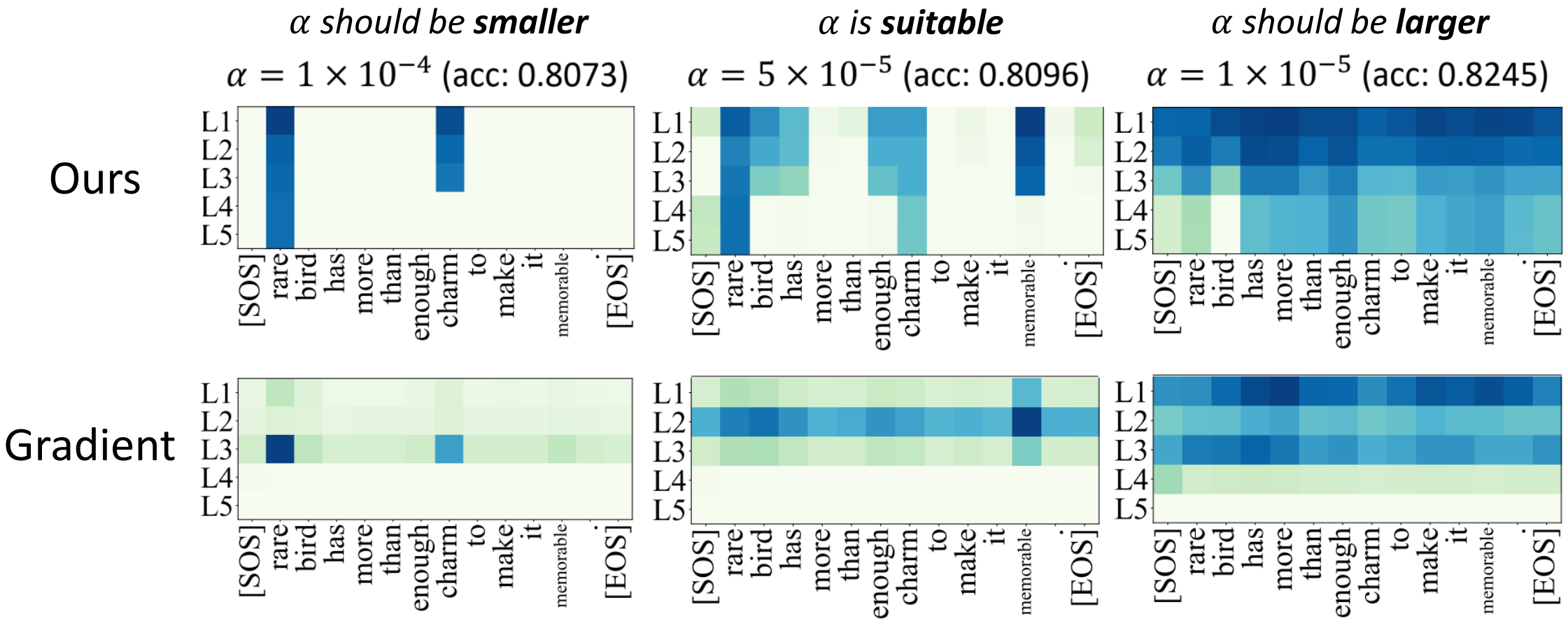






Coherency Across Models

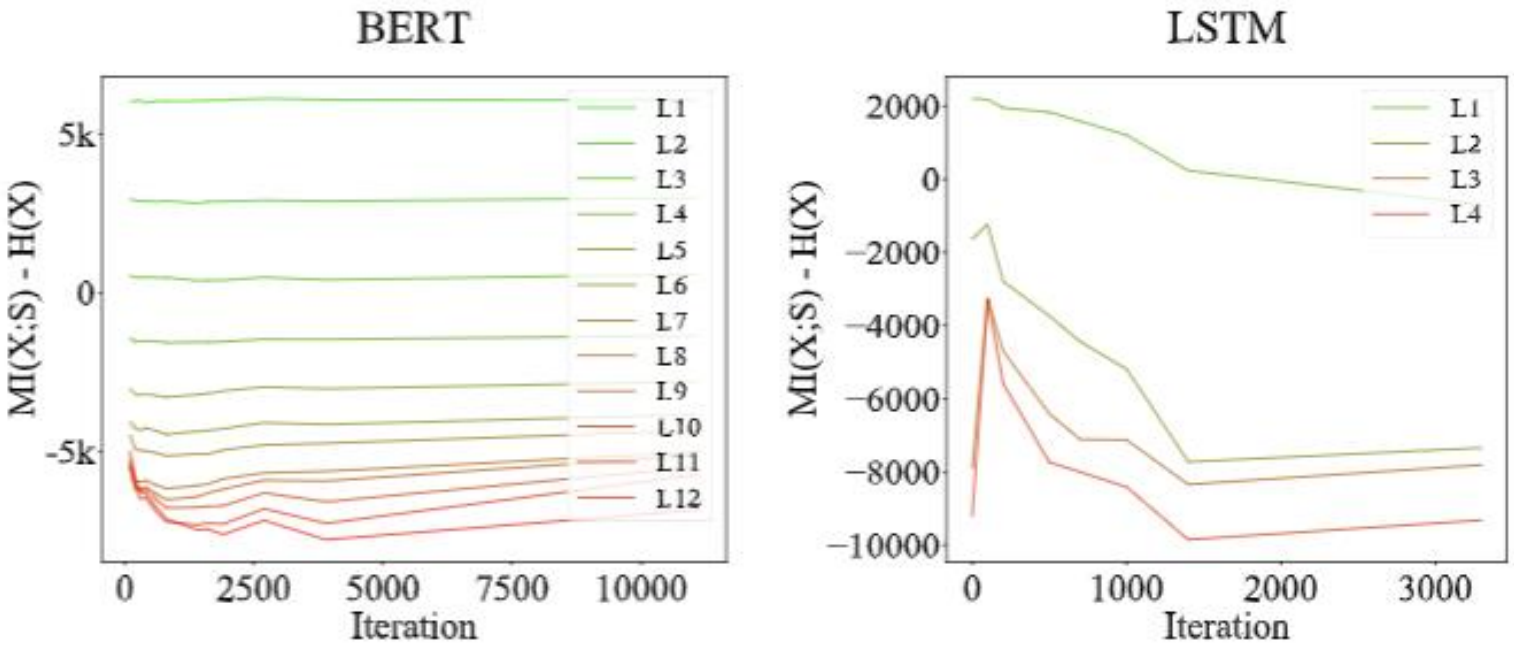
Our method provides a unified comparison of models
This enables a guided selection of regularization hyperparameter α



Understand Neural Models in NLP

How information changes during the training process

Task: sentiment classification



Conclusions:

1. For BERT, only the last few layers (L6-L12) have a considerate change of information
2. Models trained from scratch (e.g. LSTM) have information expansion at first, while pre-trained models are more stable when fine-tuning.

Impact of Unified Information Explainer

Unified Information Explainer integrated into two Azure Github repos (in total 6.4K stars)

Repo 1

microsoft / nlp-recipes Public archive — Unwatch 188 — Fork 895 — Star 6.1k

Towards a Deep and Unified Understanding of Deep Neural Models in NLP

This submodule contains a tool for explaining hidden states of models. It is an implementation of the paper *Towards a Deep and Unified Understanding of Deep Neural Models in NLP*

Repo 2

interpretml / interpret-text Public — Watch 19 — Fork 65 — Star 364

	Classical Text Explainer	Unified Information Explainer	Introspective Rationale Explainer
Input model support	Scikit-learn linear models and tree-based models	PyTorch	PyTorch
Explain BERT	No	Yes	Yes

Towards a Deep and Unified Understanding of Deep Neural Models in NLP, ICML 2019

Impact of Unified Information Explainer

Ivan Titov, Program Co-chair at ICLR 2021, Action editor for JMLR and TACL

noise to these messages. Therefore, GRAPHMASK can be categorised as belonging to the recently introduced class of perturbation-based methods (Guan et al., 2019; Taghanaki et al., 2019; Schulz et al., 2020) which equate feature importance with sensitivity of the prediction to the perturbations

Prof. Titov said his proposed method GraphMask **falls into the same class of our method**

Ge Wang, IEEE Fellow, Clark & Crossan Endowed Chair Professor, Google Scholar Citation 36k

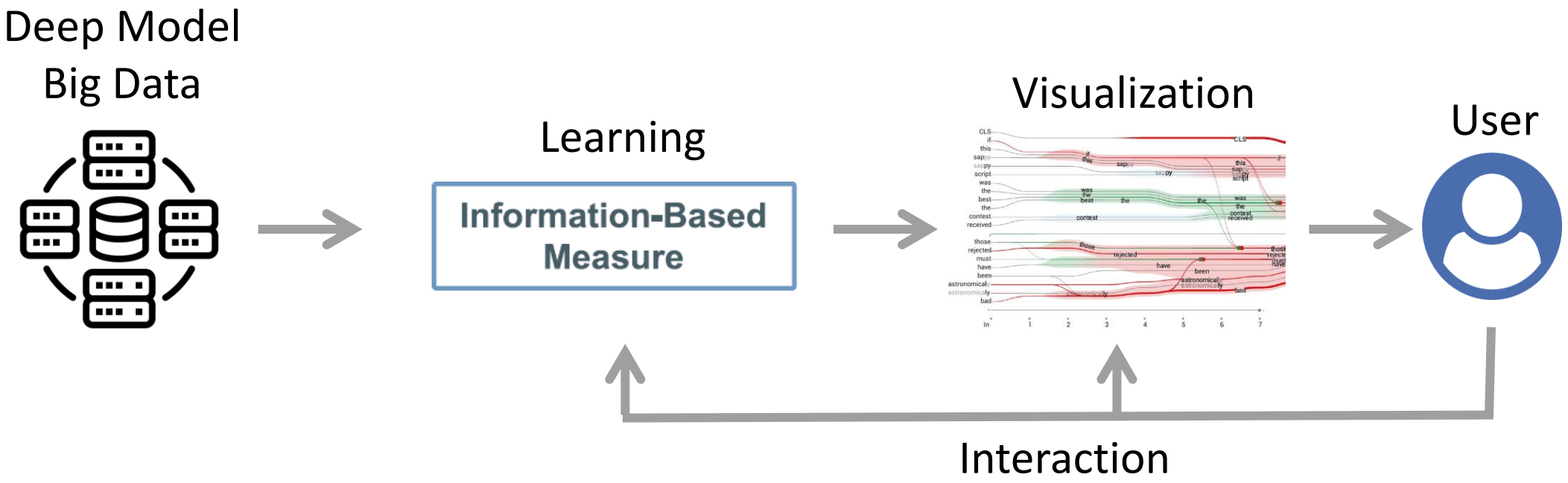
region and smallest destroying region. Guan *et al.* (2019) proposed to use mutual-information measure to quantify the association between inputs and latent representations of a deep model for natural language processing, which is coherent and general. Due to the difficulty in computing the mutual information directly, they approximated the mutual information measure by perturbation with a known distribution.

Prof. Wang said our method is **coherent and general**, and **Prof. Augenstein** highlights that our method can analyze **multiple models**

Isabelle Augenstein, Head of the Copenhagen NLU research group

(Lertvittayakumjorn and Toni, 2019), whereas a few consider **more than one** model (Guan et al., 2019; Poerner et al., 2018). Some studies concen-

Empowering Model Interpretation with Visual Analytics



Analyzing Richer Information

Inter-Word Information

Learned phrases

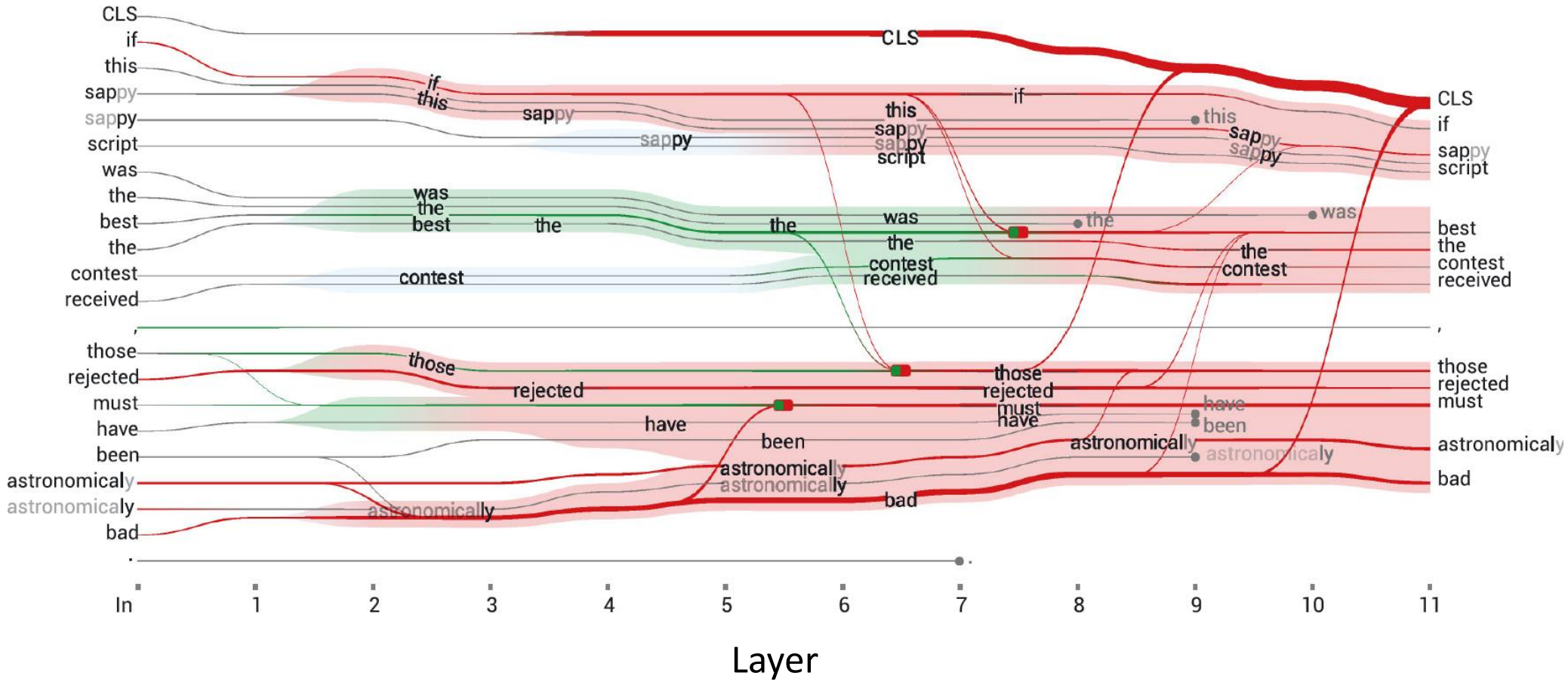
Non-adjacent word relationships

Information-Based Measure

Intra-Word Information

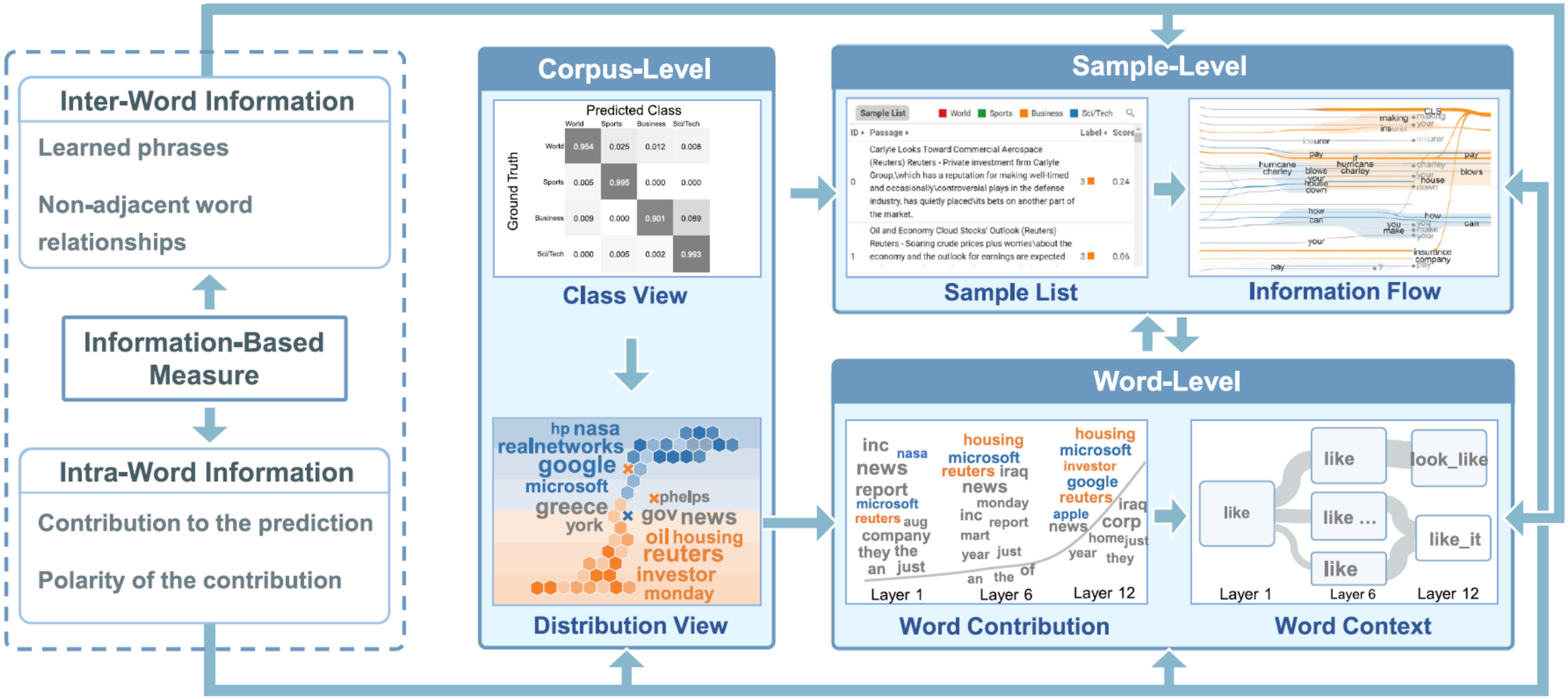
Contribution to the prediction

Polarity of the contribution



A Unified Understanding of Deep NLP Models for Text Classification, IEEE TVCG 2022

Guided Exploration through Multiple Granularities



A Unified Understanding of Deep NLP Models for Text Classification, IEEE TVCG 2022

Case 1. Diagnose Binary Sentiment Classification

Dataset: SST-2 (Stanford sentiment treebank)
Train: 67349 samples, Test: 1821 samples

Model: 12-layer BERT, Test accuracy: 93.23%

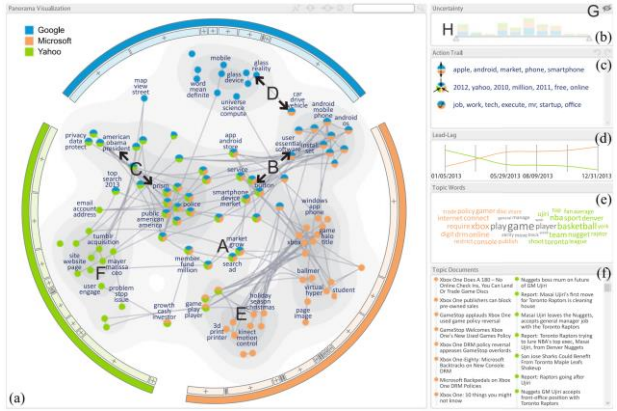
In this case, we demonstrate how DeepNLPVis helps understand and diagnose BERT for sentiment classification.

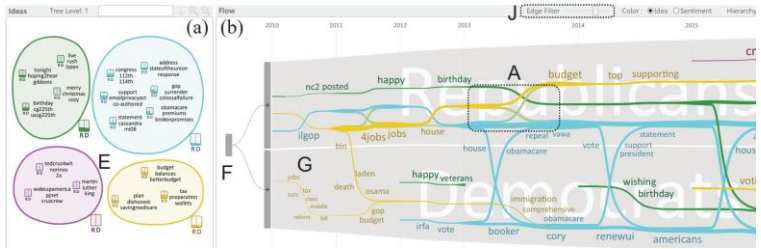
Visual Analytics of Various Models

2013 2018 2023

Classic Models → **Deep Models**

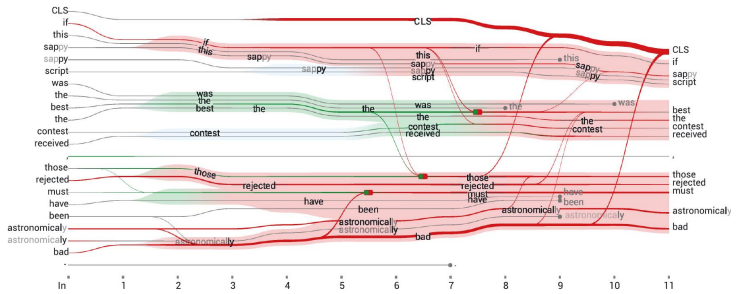
Topic models



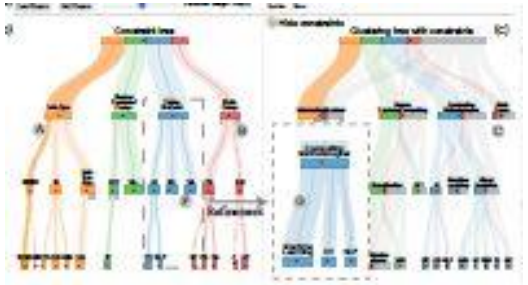


8 IEEE TVCG papers
2 IEEE VIS papers

NLP models



Word embeddings

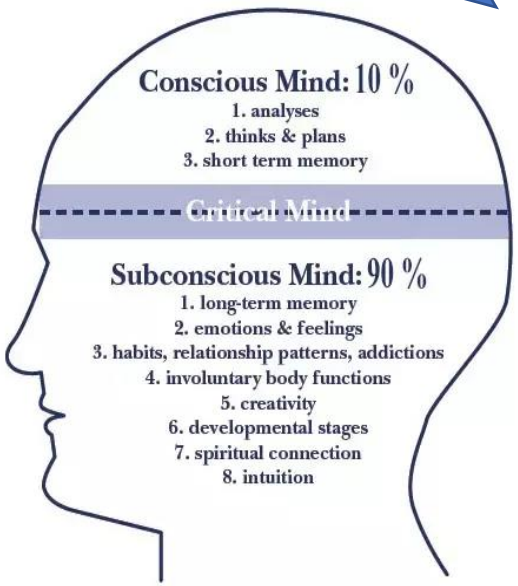


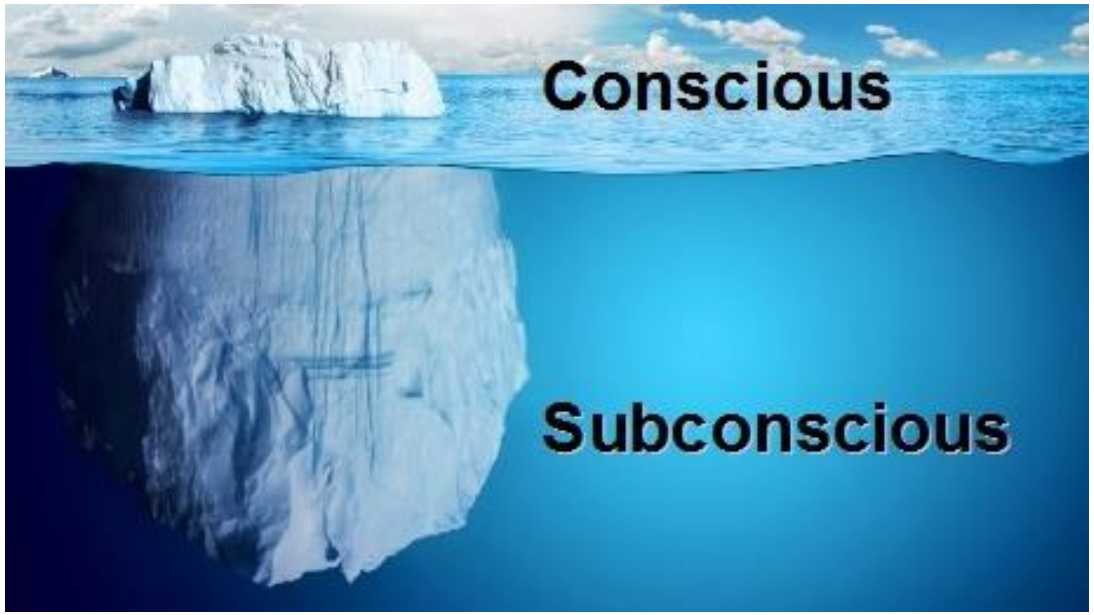
Self-Explaining Deep Models

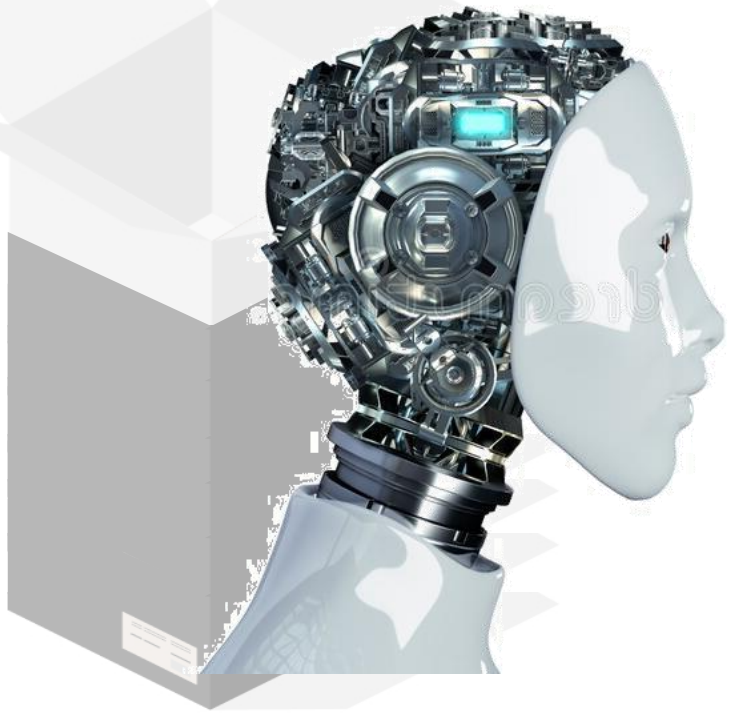
Fully transparent and easily steerable
Suitable for high-stake scenarios

Conscious Part Aligned with Humans ⇨ How to formulate this part?

No ground-truth labels





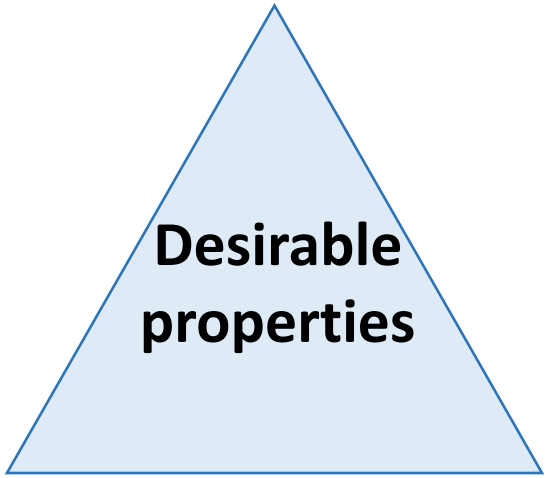


Subconscious Part Different

Explainable Part Formulation: Desirable Properties

Do not require ground-truth labels for explanations

Minimum human efforts



Widely used

Easy for everyone to understand

High human precision

Explanations are logically reasons according to humans

Human precision

Whether the explanation is logically reasonable according to humans

Low Human Precision:

movie => negative sentiment

High Human Precision:

worst => negative sentiment

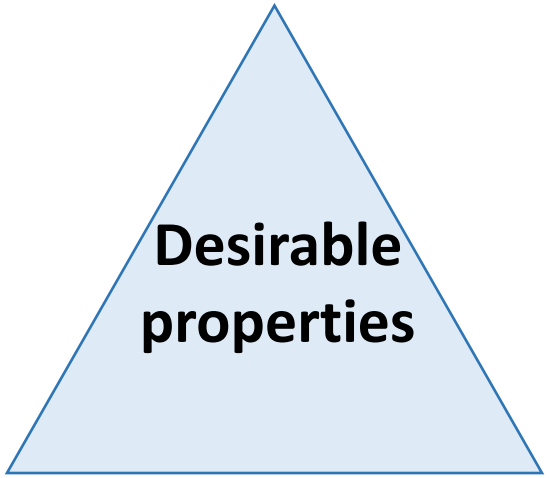
User Study shows human precision of SENN [1] is <70% (Yelp dataset)

[1] Towards Robust Interpretability with Self-Explaining Neural Networks, NeurIPS 2018

Logic Rule Satisfies All Three Desirable Properties

Do not require ground-truth labels for explanations

Minimum human efforts



Widely used

Easy for everyone to understand

High human precision

Explanations are logically reasons according to humans

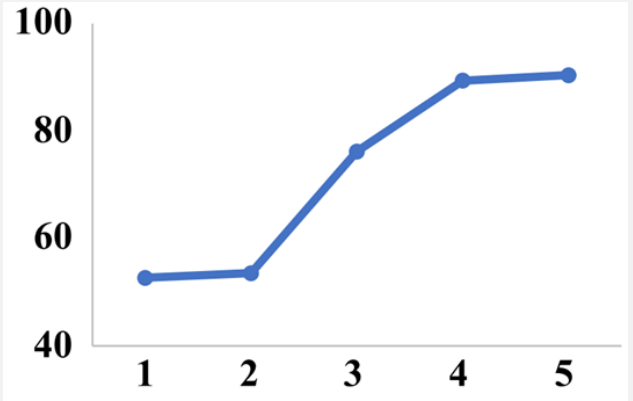
Logic rule

worst AND bad => negative sentiment (confidence from data: 95%)

Highly confident logic rules have high human precision

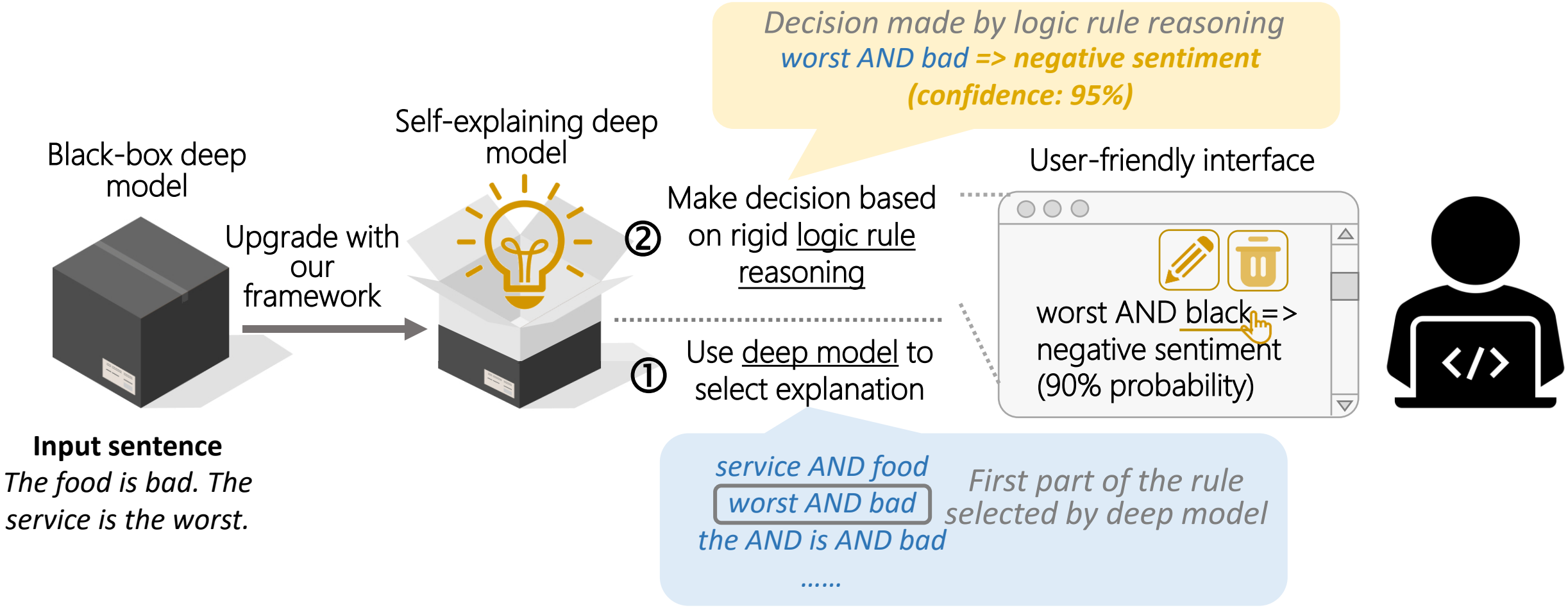
Our User Study on Yelp

Human precision



Confidence level of logic rule

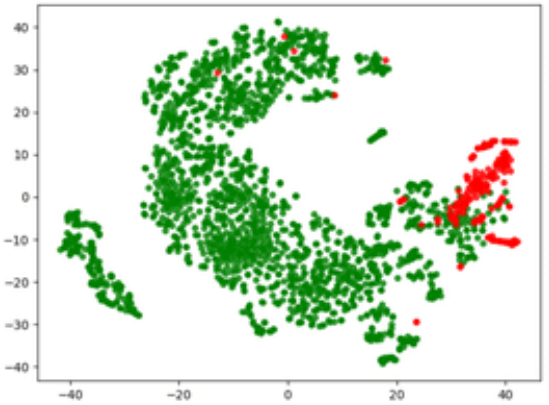
SELOR:
Self-Explaining with LOgic rule Reasoning



Self-explaining deep models with logic rule reasoning, NeurIPS 2022

Logic Rule Reasoning: Well-Grounded Decision

Reason about the decision from data



Given *"worst AND bad" selected by deep model*

1. Find all training instances that contain both *"worst"* and *"bad"*
2. 95% of them are negative
3. Model decision is *Negative Sentiment (95% confidence)*

Good property 1: Well-grounded decision

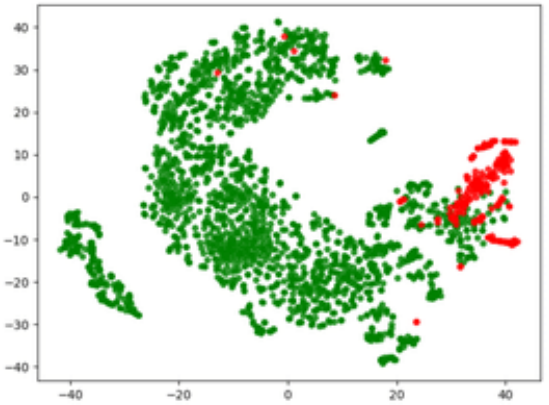
worst AND bad => Negative Sentiment (95% confidence)

is AND the => Positive Sentiment (50% confidence)

? => Positive Sentiment (99% confidence) Infeasible decision

Unifying Accuracy and Explainability

Reason about the decision from data



Given *"worst AND bad" selected by deep model*

1. Find all training instances that contain both *"worst"* and *"bad"*
2. 95% of them are negative
3. Model decision is *Negative Sentiment (95% confidence)*

Optimized with cross-entropy loss

Good property 2: Unifying the optimization of classification accuracy and explainability

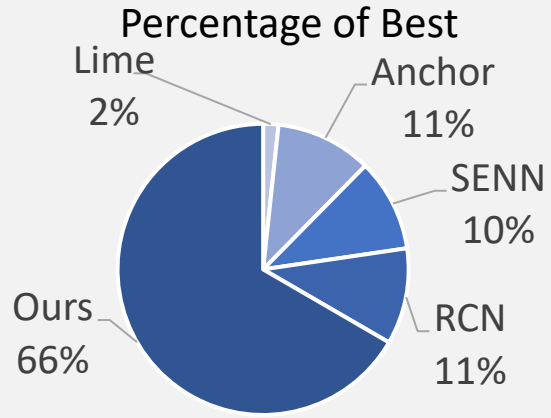
Optimizing accuracy = Maximizing logic rule confidence = Optimizing explainability

worst AND bad => Negative Sentiment (95% confidence)

is AND the => Positive Sentiment (50% confidence)

Results

High Human Precision

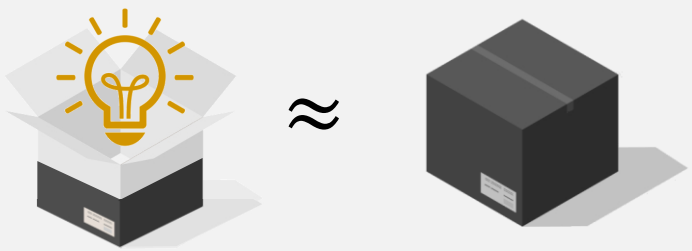


User study
Percentage of best

+500%

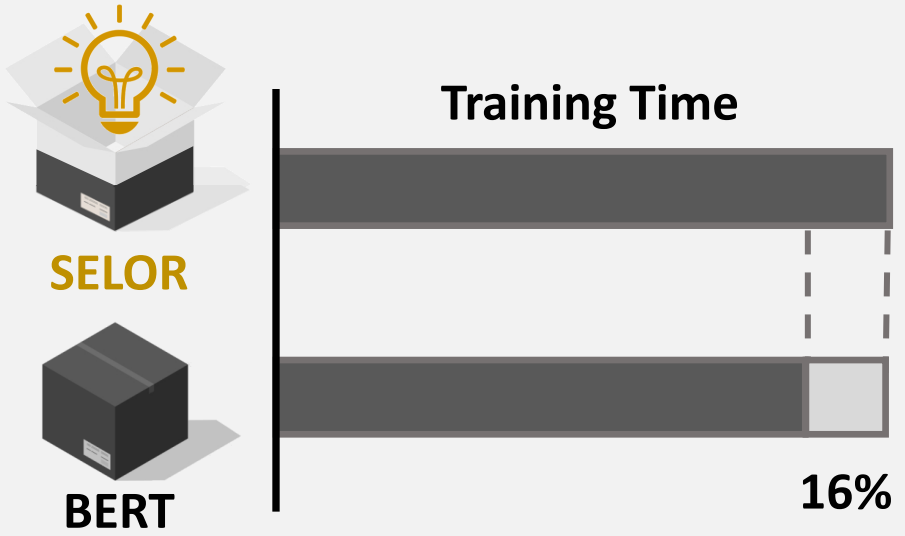
(Adult dataset)

Good Prediction Performance



Training Cost

- **Efficient, differentiable** training
- **Slightly slower** than black-box model



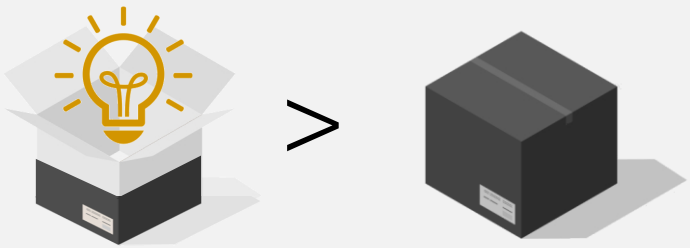
Additional Advantages

Generate Explanation Efficiently

SELOR vs LIME
1,000x speed-up

SELOR vs Anchor
50,000x speed-up
(BERT base, Yelp)

Robust to Noisy Labels

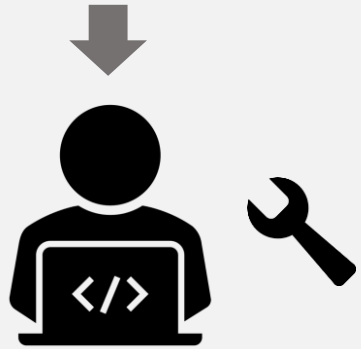


SELOR > **Black-box**

>10% F1 increase when 20% labels are randomly flipped
(BERT base, Yelp)

Can be Steered w/o Retraining

✗ *vegas* => positive



✓ *tasteless* => negative

Impact of SELOR

Well received according to audience interaction data

本篇为大家揭晓年度「最受欢迎讲者」和「人气实验室」榜单。我们对过去一年中来到社区分享工作的嘉宾，按照其Talk观看量、互动量、新媒体传播度等综合维度进行了人气排位，最终评选出25位讲者成为「2023年度TechBeat红人榜——最受欢迎讲者」，他们凭借最新

王希廷	微软亚洲研究院 高级研究员	基于逻辑规则推理的 深度自可解释模型

Microsoft list of notable research papers from 2022

Office of the Chief Technology Officer

2022 Research Reading List ☆

Paper Title	Source	Subject Ar...
Self-Explaining Deep models with Logic Rule Reasoning	arXiv	Responsible AI

2022 Research Reading List

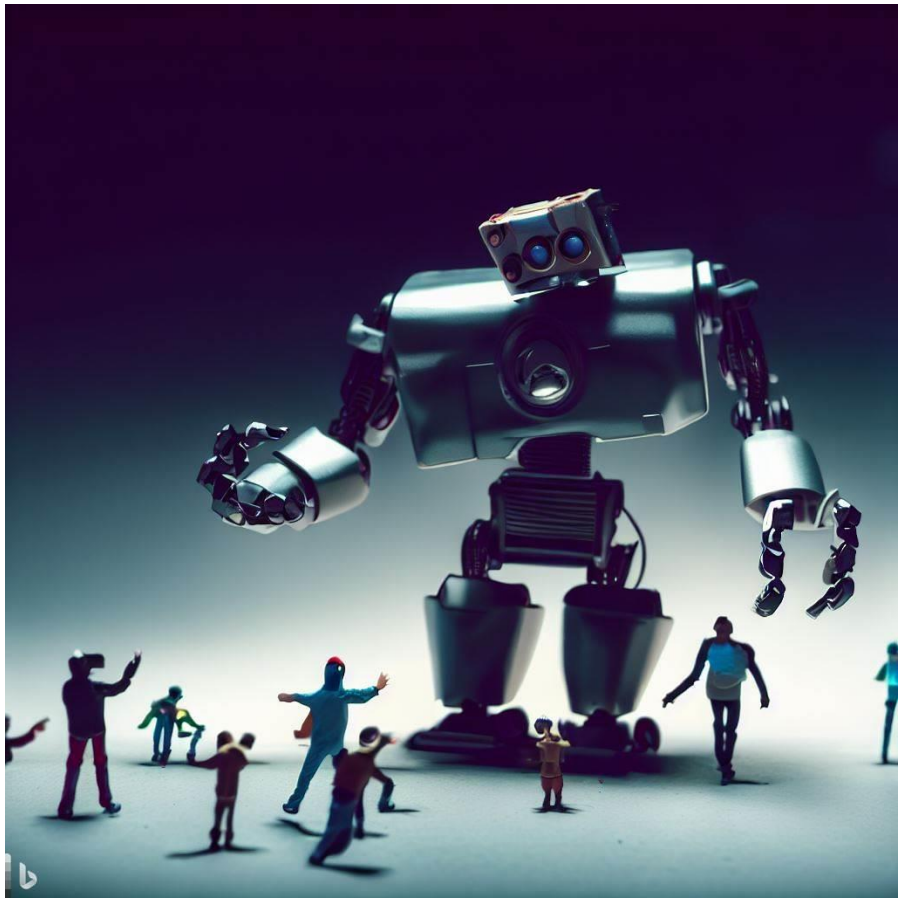
This reading list of notable research papers from 2022 was compiled by the Office of the CTO from recommendations provided by the Technical Leader Community and others. Feel

Future with Large Models

Model interpretation and alignment have become even more important in the era of large models

They are two of the seven research directions sponsored by OpenAI:

- ☐ Interpretability / Transparency: How do these models work, mechanistically? Can we identify what concepts they're using, or extract latent knowledge from the model, make inferences about the training procedure, or predict surprising future behavior?
- ☐ Alignment: How can we understand what objective, if any, a model is best understood as pursuing? How do we increase the extent to which that objective is aligned with human preferences, such as via prompt design or fine-tuning?



Digital art created by new Bing