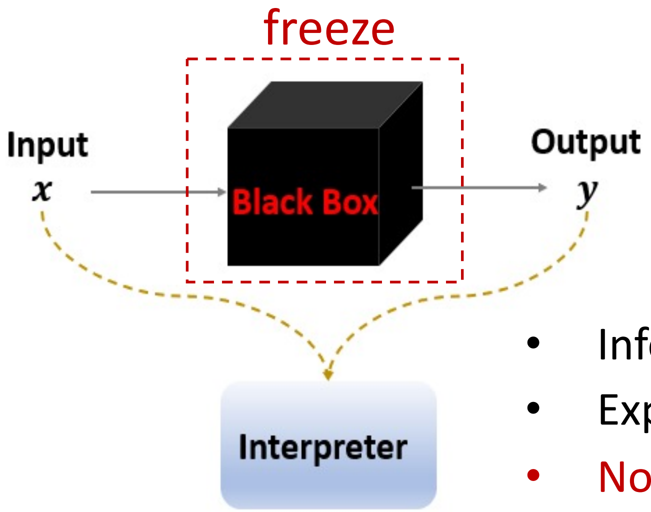
CS 4501/6501 Interpretable Machine Learning

Building Interpretable Neural Network Models

Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

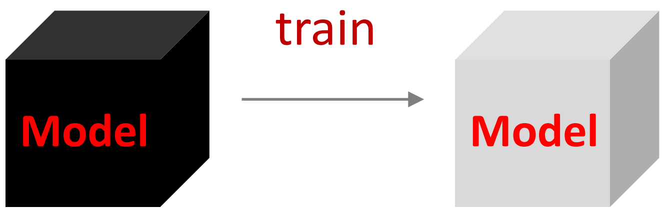
What is the difference?

Explaining a model from the post-hoc manner



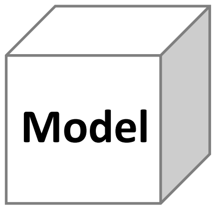
- Inference stage
- Explain model predictions
- No change on model decision making

Improving a model's intrinsic interpretability



- Training stage
- Make model prediction behavior more interpretable
- No (or minor) change on model architecture

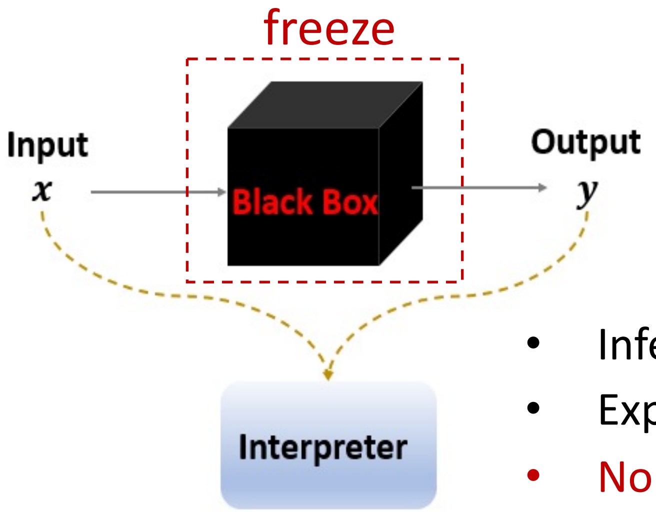
Building Interpretable Neural Network Models



Self-interpretable

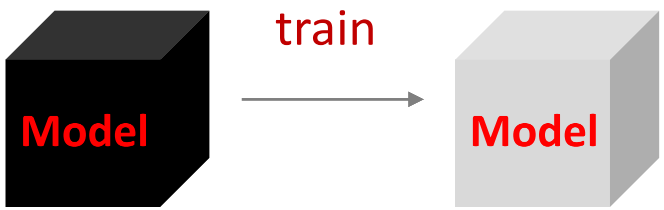
What is the difference?

Explaining a model from the post-hoc manner



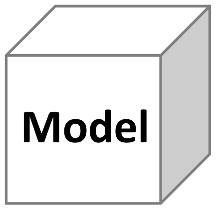
- Inference stage
- Explain model predictions
- No change on model decision making

Improving a model's intrinsic interpretability

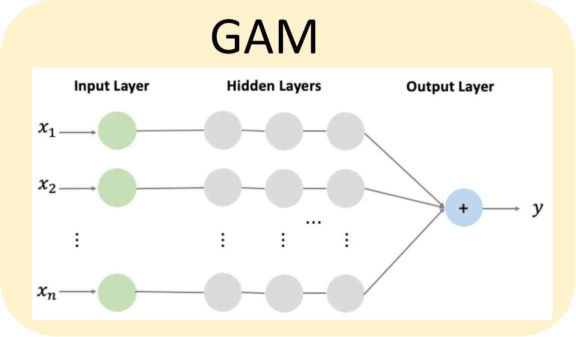


- Training stage
- Make model prediction behavior more interpretable
- No (or minor) change on model architecture

Building Interpretable Neural Network Models

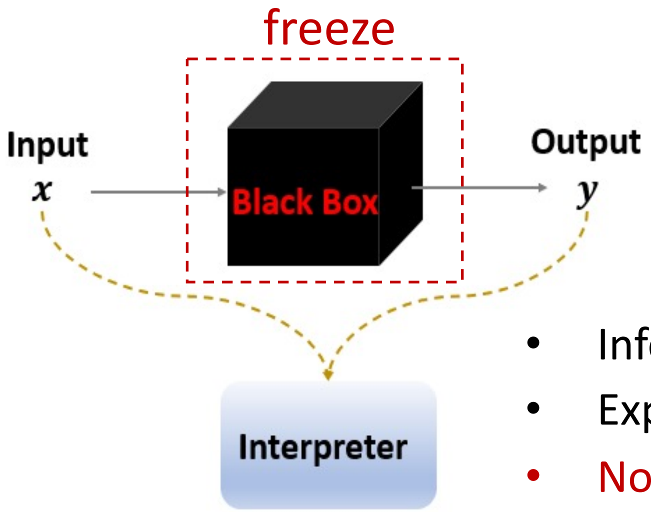


Self-interpretable



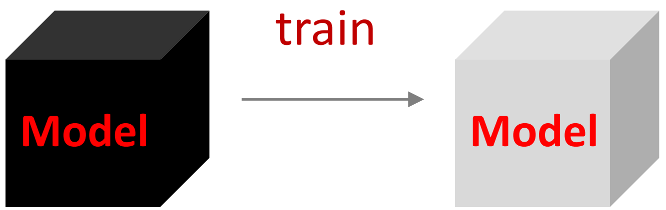
What is the difference?

Explaining a model from the post-hoc manner



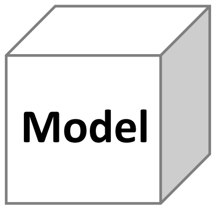
- Inference stage
- Explain model predictions
- No change on model decision making

Improving a model's intrinsic interpretability



- Training stage
- Make model prediction behavior more interpretable
- No (or minor) change on model architecture

Building Interpretable Neural Network Models



Self-interpretable

✓ Comparable or better performance to traditional neural networks

Building Interpretable Neural Networks

- Self-explaining models

- SELFEXPLAIN

Towards Robust Interpretability with Self-Explaining Neural Networks

David Alvarez-Melis, Tommi S. Jaakkola

(NeurIPS, 2018)

Goal

Building complex self-explaining models

- Providing human-interpretable explanations
- Maintaining competitive performance

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Generalized coefficients

$$f(x) = \theta(x)^T x$$

$\theta \in \Theta$ (a complex model class)

As powerful as any deep neural network, but not interpretable

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$ Feature contribution $\{\theta_i\}$

Generalized coefficients

$$f(x) = \theta(x)^T x$$ $\theta \in \Theta$ (a complex model class)

Local interpretability

$$x \approx x' \qquad \theta(x) \approx \theta(x')$$

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Generalized coefficients

$$f(x) = \theta(x)^T x$$

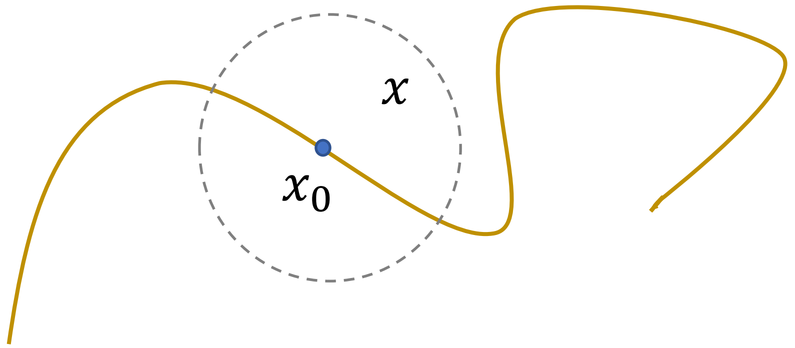
$\theta \in \Theta$ (a complex model class)

Local interpretability

$$x \approx x' \qquad \theta(x) \approx \theta(x')$$

$$\nabla_x f(x) \approx \theta(x_0)$$

The stable coefficients $\{\theta(x_0)_i\}$ indicate feature importance in the local area



Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Beyond raw features – feature basis

Interpretable basis concepts: higher order features (e.g., a patch of pixels)

$$h(x)\colon \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$$

(k is small for interpretation)

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Beyond raw features – feature basis

Interpretable basis concepts: higher order features (e.g., a patch of pixels)

$$h(x)\colon \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$$

(k is small for interpretation)

- subset aggregates of the input (e.g., $h(x) = Ax$, A is a boolean mask matrix)
- predefined, pre-grounded feature extractors designed with expert knowledge (e.g., filters for image processing)
- prototype based concepts

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Beyond raw features – feature basis

Interpretable basis concepts: higher order features (e.g., a patch of pixels)

$$h(x)\colon \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$$

(k is small for interpretation)

$$f(x) = \theta(x)^T h(x) = \sum_{i=1}^{k} \underline{\theta(x)_i} h(x)_i$$

Concept importance

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Further generalization

$$f(x) = \theta(x)^T h(x) = \sum_{i=1}^{k} \theta(x)_i h(x)_i$$

$\sum \longrightarrow g(z_1, \cdots, z_k)$ (a general aggregation function)

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Further generalization

$$f(x) = \theta(x)^T h(x) = \sum_{i=1}^{k} \theta(x)_i h(x)_i$$

$\sum \longrightarrow g(z_1, \cdots, z_k)$ (a general aggregation function)

- be permutation invariant
- isolate the effect of individual $h(x)_i$ in the output
- preserve the sign and relative magnitude of the impact of the relevance values $\theta(x)_i$

Interpretability: linear and beyond

Linear regression

$$f(x) = \sum_{i=1}^{n} \theta_i x_i$$

Feature contribution $\{\theta_i\}$

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

$g(\cdot)$: aggregation function

$h(x)$: basis concepts

$\theta \in \Theta$: a complex model

(conditional bounding $\|\theta(x) - \theta(y)\|$ with $L\|h(x) - h(y)\|$)

θ acts as coefficients of a linear model on the basis concepts $h(x)$

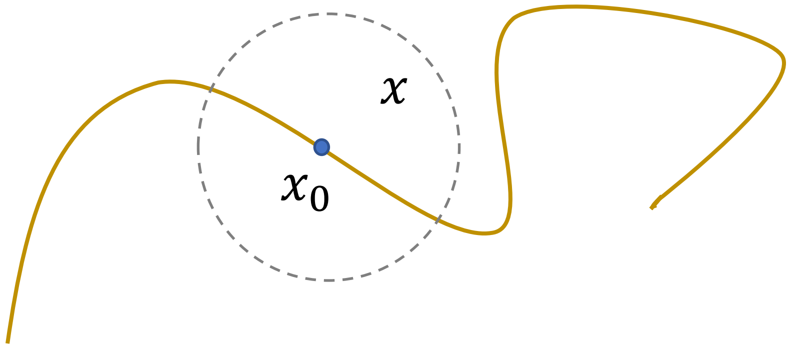
Question?

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

For every x_0, there exist $\delta > 0$ and $L \in \mathbb{R}$ such that $\|x - x_0\| < \delta$ implies $\|\theta(x) - \theta(x_0)\| \leq L\|h(x) - h(x_0)\|$



Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \big\{\big(h_i(x), \theta_i(x)\big)\big\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

$h(\cdot)$ is a trivial input feature indicator, while the modeling capacity comes from $\theta(\cdot)$ (e.g., DNNs)

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \big\{\big(h_i(x), \theta_i(x)\big)\big\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$

$\sum z_i$ or $\sum A_i z_i$ $(A_i > 0)$

- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \big\{\big(h_i(x), \theta_i(x)\big)\big\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

Application-dependent

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \big\{\big(h_i(x), \theta_i(x)\big)\big\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

$\theta(x_0) \approx \nabla_z f$

$z = h(x)$ (around x_0)

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \{(h_i(x), \theta_i(x))\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

$\theta(x_0) \approx \nabla_z f$

$z = h(x)$ (around x_0)

$\nabla_x f = \nabla_z f \underbrace{J_x^h}_{\text{(Jacobian)}}$ (chain rule)

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \big\{\big(h_i(x), \theta_i(x)\big)\big\}_{i=1}^{k}$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- $\boxed{\theta \text{ is locally difference bounded by } h}$
- $h(x)$ is an interpretable representation of x
- k is small

$\theta(x_0) \approx \nabla_z f$

$z = h(x)$ (around x_0)

$\nabla_x f = \nabla_z f J_x^h$ (chain rule)

$\theta(x)^T J_x^h \approx \nabla_x f$

$\mathcal{L}_\theta(f(x)) = \|\nabla_x f(x) - \theta(x)^T J_x^h(x)\| \approx 0$

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \{(h_i(x), \theta_i(x))\}_{i=1}^k$ of basis concepts and their influence scores

Self-explaining models

$$f(x) = g\big(\theta_1(x)h_1(x), \cdots, \theta_k(x)h_k(x)\big)$$

- g: monotone and completely additively separable
- For every $z_i = \theta_i(x)h_i(x)$, g satisfies $\frac{\partial g}{\partial z_i} \geq 0$
- θ is locally difference bounded by h
- $h(x)$ is an interpretable representation of x
- k is small

Objective

$\mathcal{L}_y(f(x), y) + \lambda \mathcal{L}_\theta(f(x))$

$\theta(x_0) \approx \nabla_z f$

$z = h(x)$ (around x_0)

$\nabla_x f = \nabla_z f J_x^h$ (chain rule)

$\theta(x)^T J_x^h \approx \nabla_x f$

$\mathcal{L}_\theta(f(x)) = \left\| \nabla_x f(x) - \theta(x)^T J_x^h(x) \right\| \approx 0$

The explanation of $f(x)$ is the set $\mathcal{E}_f(x) = \left\{\big(h_i(x), \theta_i(x)\big)\right\}_{i=1}^k$ of basis concepts and their influence scores

Question?

Learning interpretable basis concepts

$h(x)\colon \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$

single pixels → textures, shapes

single words → phrases

Ideally, the basis concepts would be informed by expert knowledge (e.g., doctor-provided features)

Learning interpretable basis concepts

$h(x)\colon \mathcal{X} \rightarrow \mathcal{Z} \subset \mathbb{R}^k$c

single pixels → textures, shapes

single words → phrases

Learning $\boldsymbol{h}$

- training h as an autoencoder
- enforcing diversity through sparsity (few non-overlapping concepts)
- providing interpretation on the concepts by prototyping (e.g., by providing a small set of training examples that maximally activate each concept)

$$\mathcal{L}_h(x, \hat{x})$$

$$\hat{x} = h_{dec}(h(x))$$

(reconstruction)

Learning interpretable basis concepts

$h(x)\colon \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$c

single pixels → textures, shapes

single words → phrases

Learning $\boldsymbol{h}$

- training h as an autoencoder
- enforcing diversity through sparsity (few non-overlapping concepts)
- providing interpretation on the concepts by prototyping (e.g., by providing a small set of training examples that maximally activate each concept)

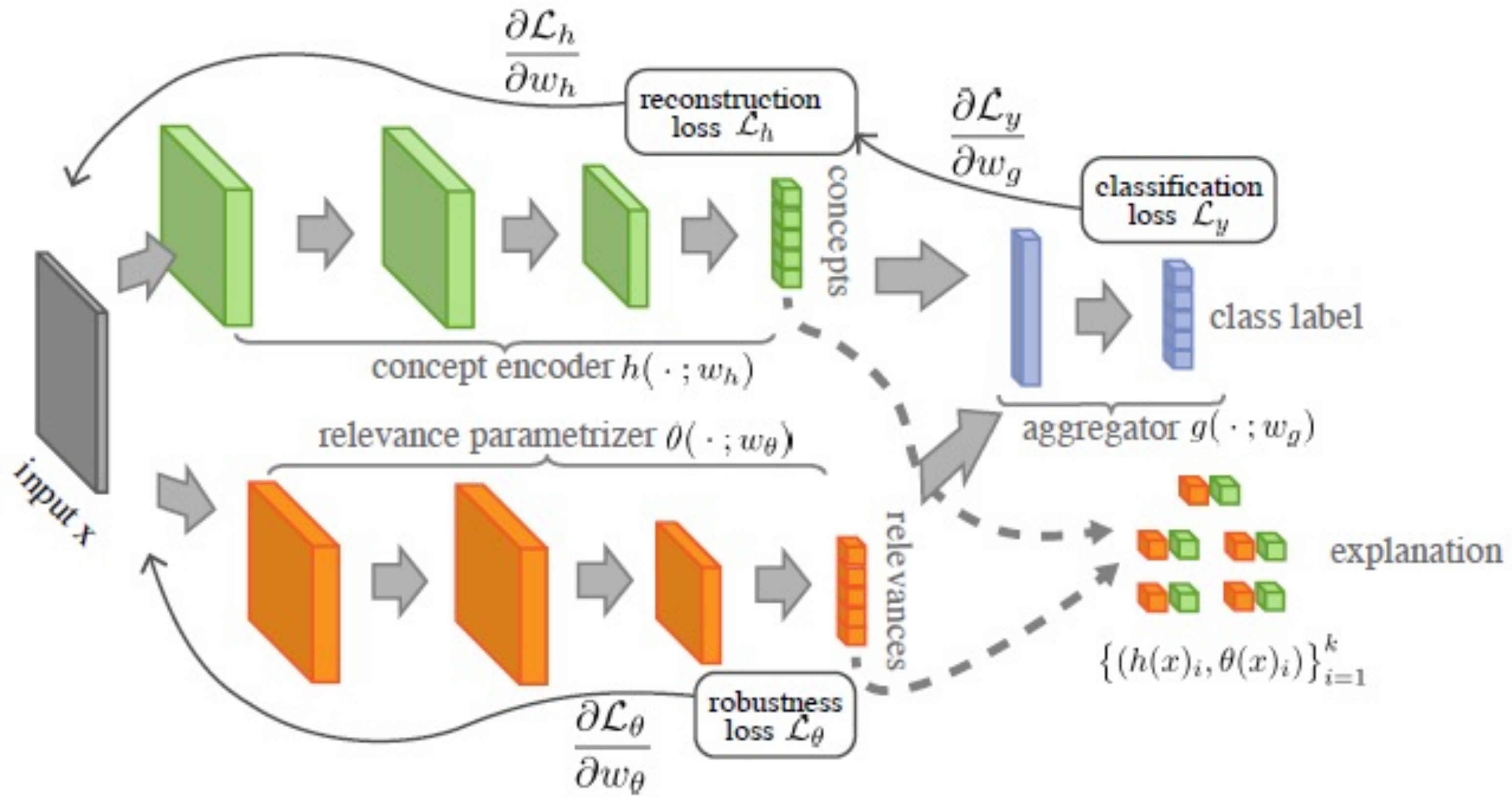
$$\mathcal{L}_h(x, \hat{x})$$

$$\hat{x} = h_{dec}(h(x))$$

Objective

$$\mathcal{L}_y(f(x), y) + \lambda \mathcal{L}_\theta(f(x)) + \gamma \mathcal{L}_h(x, \hat{x})$$

Learning interpretable basis concepts



Objective

$$\mathcal{L}_y(f(x), y) + \lambda \mathcal{L}_\theta(f(x)) + \gamma \mathcal{L}_h(x, \hat{x})$$

Architectures

- CL: convolutional layers
- FC: fully-connected layers

	COMPAS/UCI	MNIST	CIFAR10
$h(\cdot)$	$h(x) = x$	$\mathrm{CL}(10, 20) \to \mathrm{FC}(c)$	$\mathrm{CL}(10, 20) \to \mathrm{FC}(c)$
$\theta(\cdot)$	$\mathrm{FC}(10, 5, 5, 1)$	$\mathrm{CL}(10, 20) \to \mathrm{FC}(c \cdot 10)$	$\mathrm{CL}(2^6, 2^7, 2^8, 2^9, 2^9) \to \mathrm{FC}(2^8, 2^7, c \cdot 10)$
$g(\cdot)$	sum	sum	sum

Prediction performance is comparable to baseline NNs

Question?

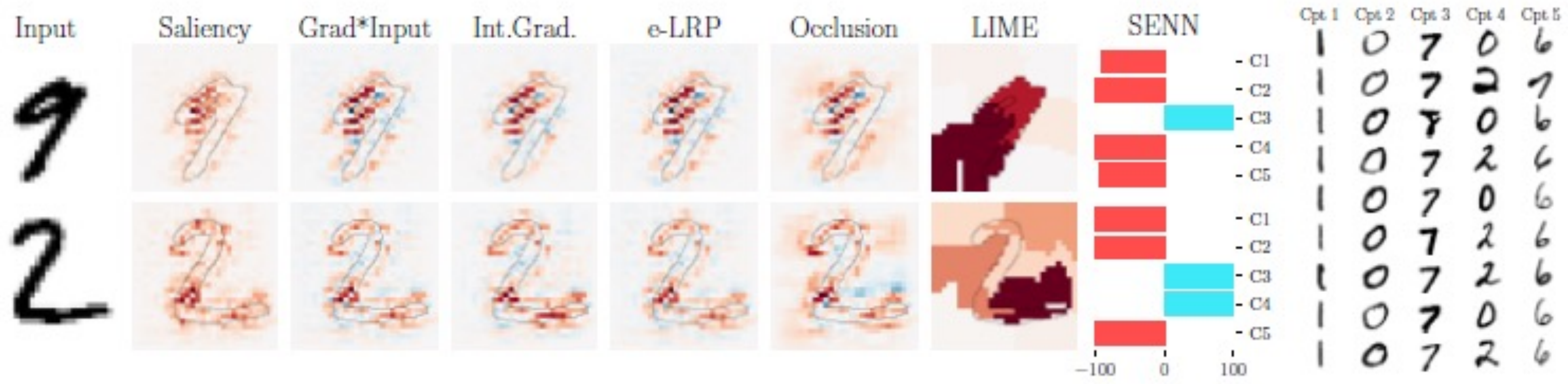
Experiments

- Explicitness/Intelligibility: Are the explanations immediate and understandable?
- Faithfulness: Are relevance scores indicative of "true" importance?
- Stability: How consistent are the explanations for similar/neighboring examples?

Experiments

Explicitness/Intelligibility: Are the explanations immediate and understandable?

- The concepts are maximally activated by a set of training examples
- Concept 3 has a strong positive influence towards both prediction
- Concept 4 is also highly relevant to “2”



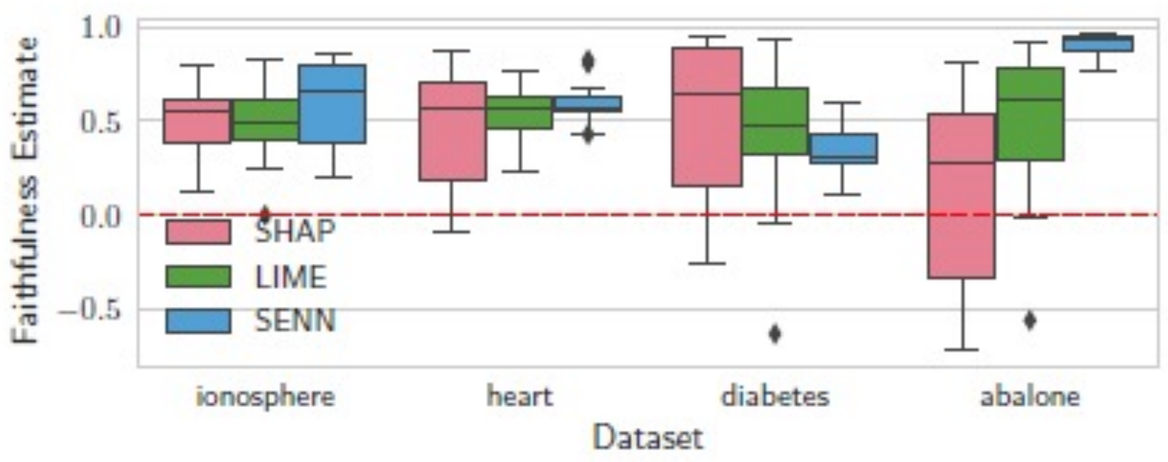
Experiments

- Explicitness/Intelligibility: Are the explanations immediate and understandable?
- Faithfulness: Are relevance scores indicative of "true" importance?
- Stability: How consistent are the explanations for similar/neighboring examples?

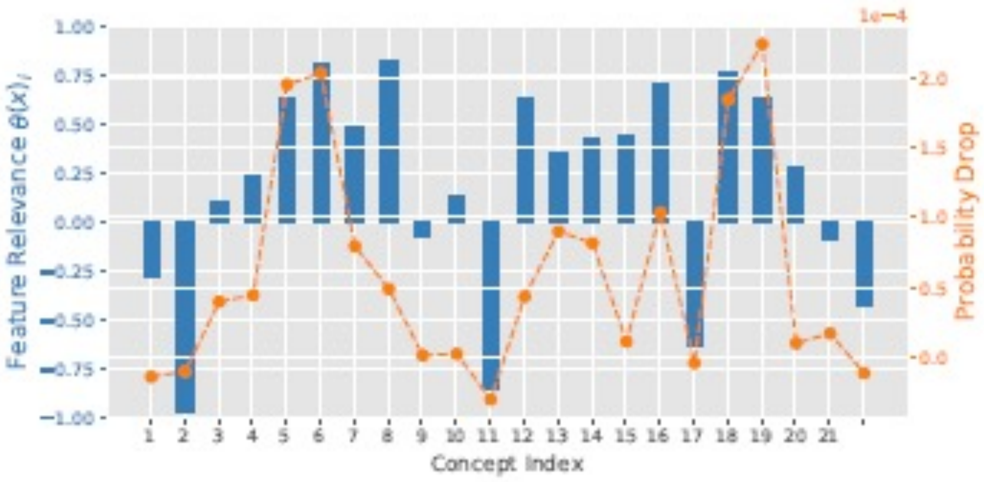
Experiments

Faithfulness: Are relevance scores indicative of "true" importance?

- Faithfulness: computing the correlations of probability drops (removing features) and relevance scores
- Overall SENN (self-explaining neural networks) can provide faithful interpretations



$h(x)$ is identity



$h(x)$ is learnt

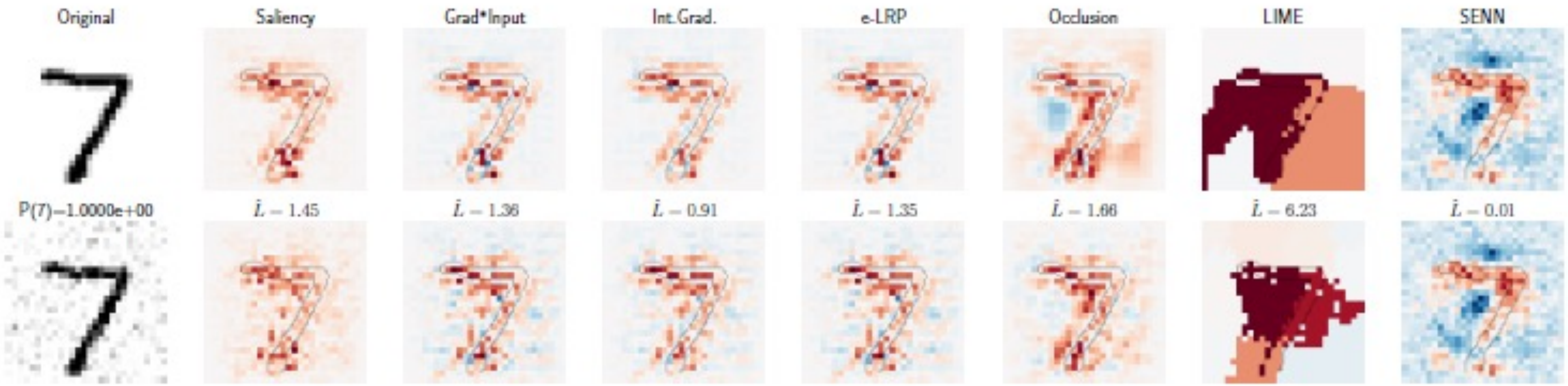
Experiments

- Explicitness/Intelligibility: Are the explanations immediate and understandable?
- Faithfulness: Are relevance scores indicative of "true" importance?
- Stability: How consistent are the explanations for similar/neighboring examples?

Experiments

Stability: How consistent are the explanations for similar/neighboring examples?

- Existing interpretation methods are not robust to small perturbations



Discussion

- Providing insights on designing self-explaining neural network models
- Model architectures are selected empirically (requiring engineering effort)
- It is still challenging to develop interpretable models in more complex domains (e.g., larger image datasets, NLP tasks)

Building Interpretable Neural Networks

- Self-explaining models

- SELFEXPLAIN

SELFEXPLAIN

SELFEXPLAIN: A Self-Explaining Architecture for

Neural Text Classifiers

Dheeraj Rajagopal, Vidhisha Balachandran,
Eduard Hovy, Yulia Tsvetkov

(EMNLP, 2021)

SELFEXPLAIN

- Local interpretable layer (LIL)

  Identifying local feature attributions in the input

- Global interpretable layer (GIL)

  Explaining model decisions as a function of influential training data

- High-level phrase-based concepts

SELFEXPLAIN

- Local interpretable layer (LIL)

  Identifying local feature attributions in the input

- Global interpretable layer (GIL)

  Explaining model decisions as a function of influential training data

- High-level phrase-based concepts

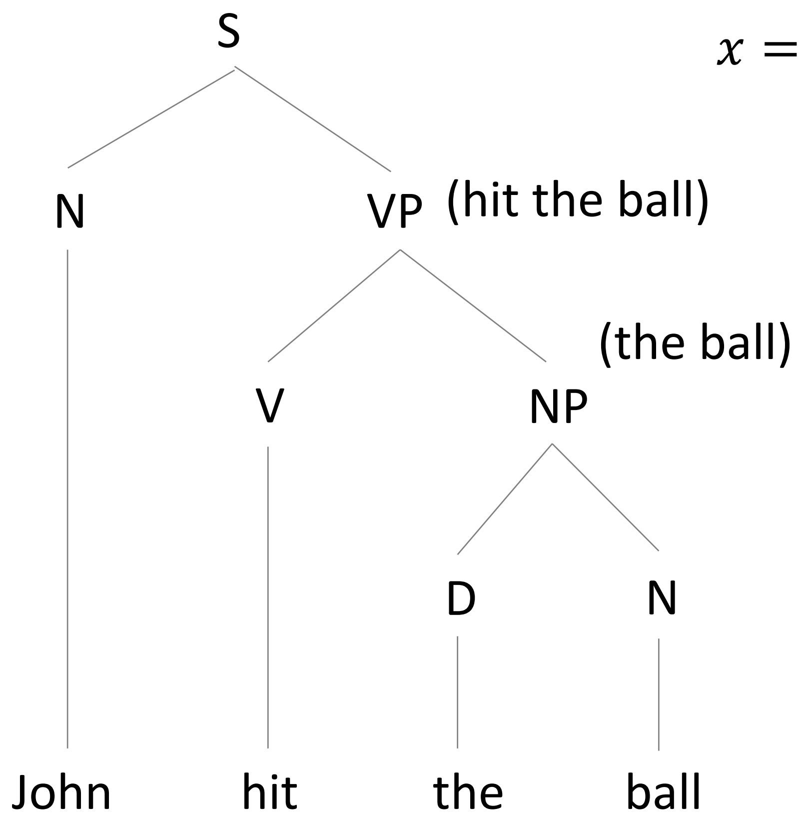
$\mathcal{M}$: a neural C-class classification model

SELFEXPLAIN builds into $\mathcal{M}$ and provides a set of explanations Z

SELFEXPLAIN

Defining human-interpretable concepts (phrases)

Extract phrases via syntax trees



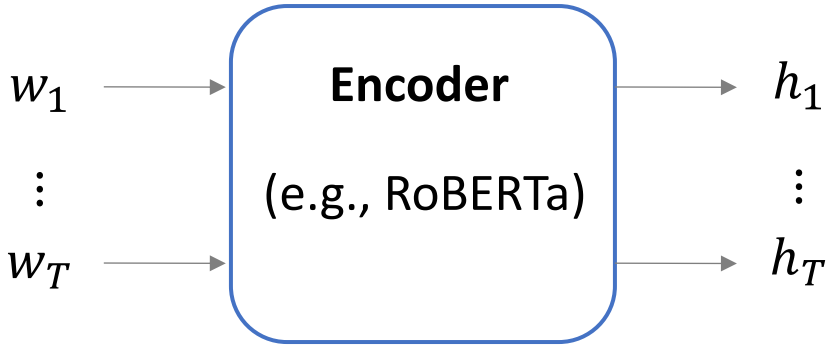
Input $\rightarrow$ Component non-terminals

$$x = \{w_i\}_{1:T} \longrightarrow N(x) = \{nt_j\}_{1:J}$$

S: sentence
NP: noun phrase
VP: verb phrase
V: verb
D: determiner
N: noun

SELFEXPLAIN

Concept-aware encoder E

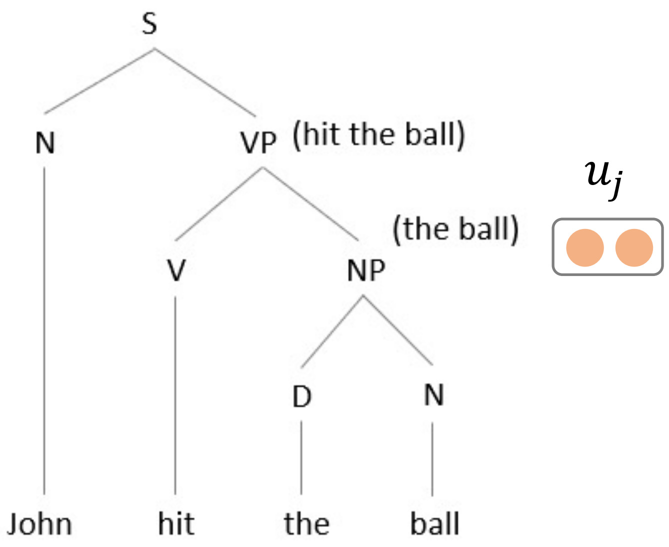


Input $\quad$ Component non-terminals

$$x = \{w_i\}_{1:T} \longrightarrow N(x) = \{nt_j\}_{1:J}$$

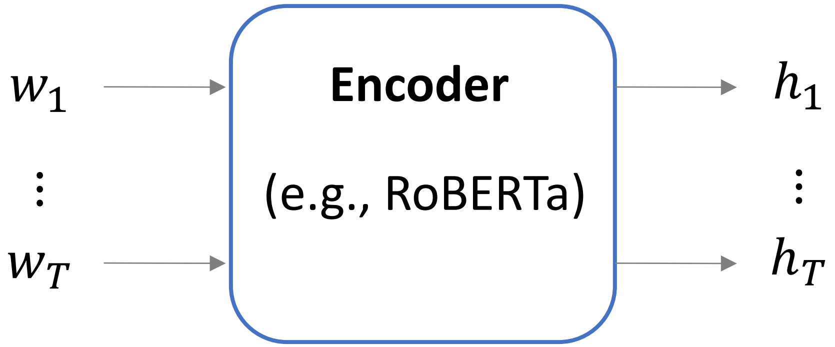
The representation of non-terminal nt_j

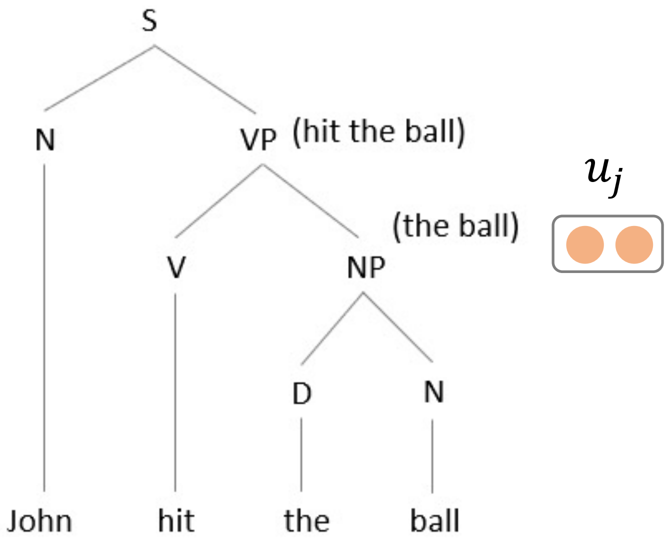
$$u_j = \frac{\sum_{w_i \in nt_j} h_i}{len(nt_j)}$$



SELFEXPLAIN

Concept-aware encoder E





Input

$$x = \{w_i\}_{1:T} \longrightarrow N(x) = \{nt_j\}_{1:J}$$

Component non-terminals

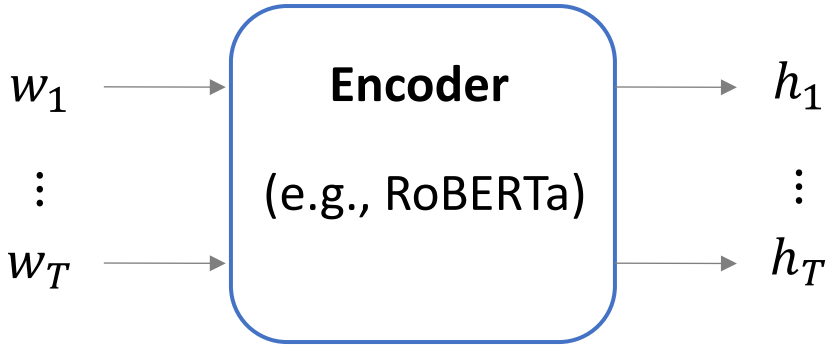
The representation of non-terminal nt_j

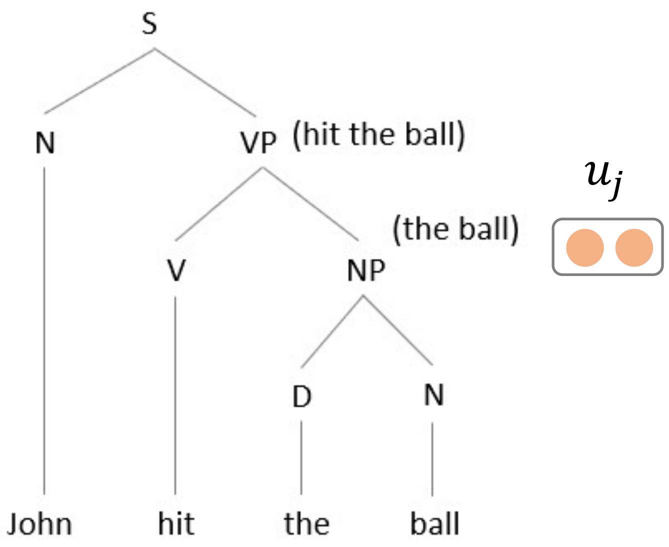
$$u_j = \frac{\sum_{w_i \in nt_j} h_i}{len(nt_j)}$$

u_S is the pooled representation ([CLS] token representation)

SELFEXPLAIN

Concept-aware encoder E





Input — Component non-terminals

$$x = \{w_i\}_{1:T} \longrightarrow N(x) = \{nt_j\}_{1:J}$$

The representation of non-terminal nt_j

$$u_j = \frac{\sum_{w_i \in nt_j} h_i}{len(nt_j)}$$

The output of the classification layer

$$l_Y = softmax\big(W_y g(u_S) + b_y\big)$$

$$P_C = argmax(l_Y)$$

$g(\cdot)$: $relu$ activation layer

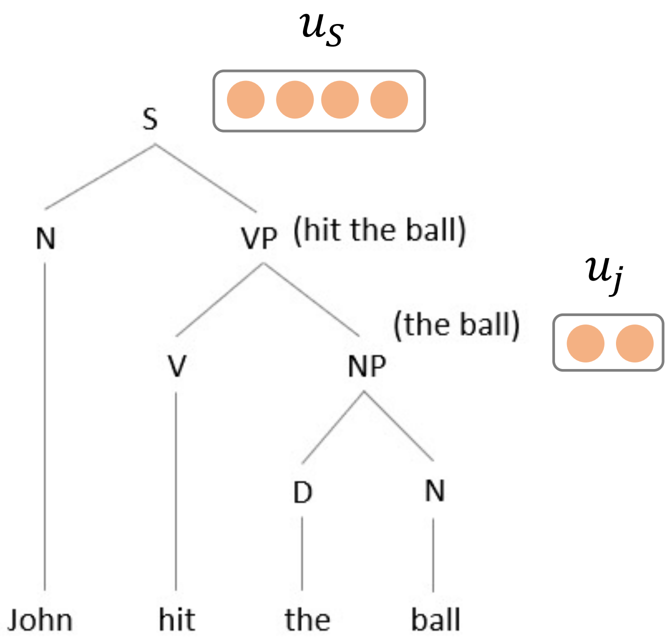
Question?

SELFEXPLAIN

Local interpretability layer (LIL)

Compute the local relevance score for all input concepts $\{nt_j\}_{1:J}$ from the sample x

Activation difference: quantifies the contribution of each nt_j to the label in comparison to the contribution of the root node nt_S

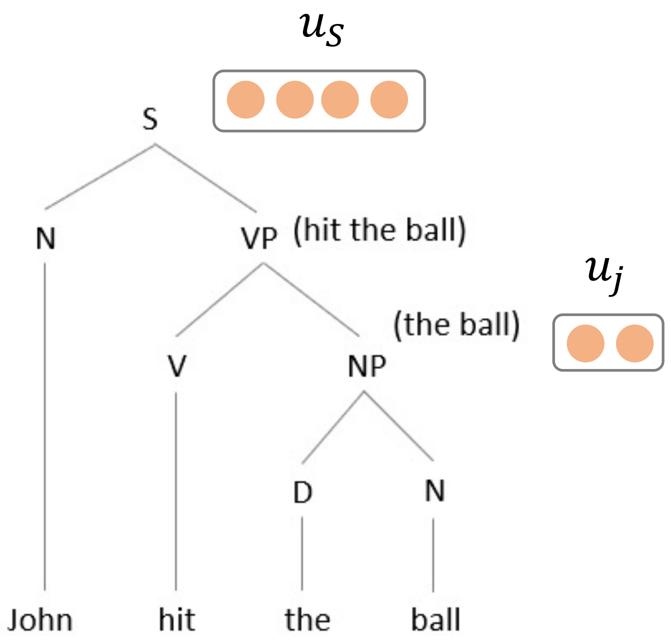


SELFEXPLAIN

Local interpretability layer (LIL)

Compute the local relevance score for all input concepts $\{nt_j\}_{1:J}$ from the sample x

Activation difference: quantifies the contribution of each nt_j to the label in comparison to the contribution of the root node nt_S



$t_j = g(u_j) - g(u_S)$ *relu* activation function

$s_j = softmax(W_v t_j + b_v)$ LIL parameters

The relevance score of nt_j

$$r_j = (l_Y)_i|_{i=P_C} - (s_j)_i|_{i=P_C}$$

Original prediction probabilities

Predicted label

Question?

SELFEXPLAIN

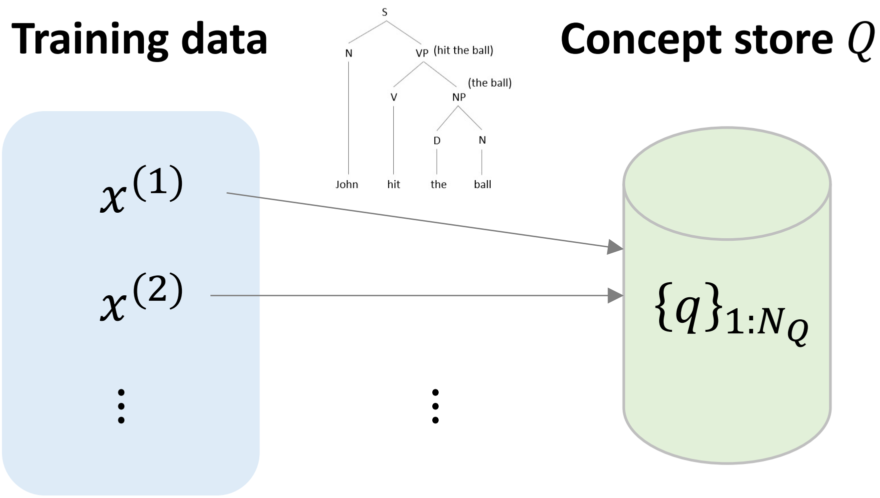
Global interpretability layer (GIL)

Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions

SELFEXPLAIN

Global interpretability layer (GIL)

Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions



$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)}$$

e: the embedding layer

SELFEXPLAIN

Global interpretability layer (GIL)

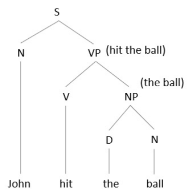
Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions

Training data

$x^{(1)}$

$x^{(2)}$

$\vdots$



$\vdots$

Concept store Q

$\{q\}_{1:N_Q}$

$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)}$$

e: the embedding layer

Retrieve K influential concepts for an input x

$\{q\}_{1:K}$

$$d(x, Q) = \frac{x \cdot q}{\|x\| \|q\|}$$

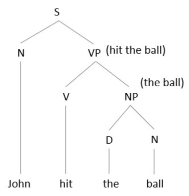
$q \in Q$

SELFEXPLAIN

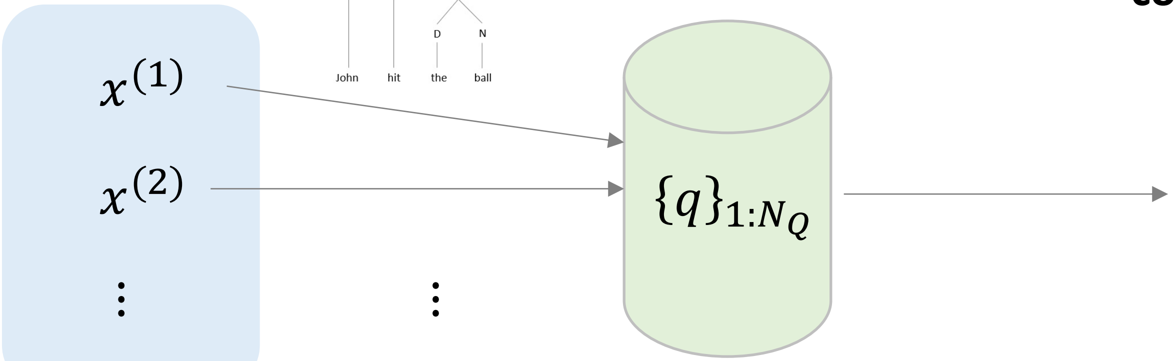
Global interpretability layer (GIL)

Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions

Training data



Concept store Q



$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)}$$

e: the embedding layer

Retrieve K influential concepts for an input x

$$\{q\}_{1:K}$$

$$d(x, Q) = \frac{x \cdot q}{\|x\|\|q\|}$$

$$q \in Q$$

Maximum Inner Product Search

$$p(q|x) = \frac{\exp(d(u_S, q))}{\sum_{q'} \exp(d(u_S, q'))}$$

SELFEXPLAIN

Global interpretability layer (GIL)

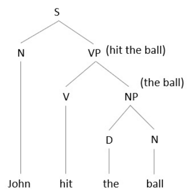
Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions

Training data

$x^{(1)}$

$x^{(2)}$

$\vdots$



$\vdots$

Concept store Q

$\{q\}_{1:N_Q}$

$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)}$$

e: the embedding layer

Retrieve K influential concepts for an input x

$\{q\}_{1:K}$

$$d(x, Q) = \frac{x \cdot q}{\|x\| \|q\|}$$

$q \in Q$

Classification

$$q_K = \sum_{k=1}^{K} w_k q_k$$

$$l_G = softmax(W_u g(q_K) + b_u)$$

SELFEXPLAIN

Global interpretability layer (GIL)

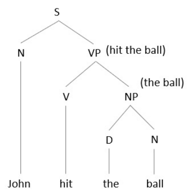
Interpret each data sample x by providing a set of K concepts from the training data which most influence the model's predictions

Training data

$x^{(1)}$

$x^{(2)}$

$\vdots$



$\vdots$

Concept store Q

$\{q\}_{1:N_Q}$

$$q_k = \frac{\sum_{w \in q_k} e(w)}{len(q_k)}$$

e: the embedding layer

Retrieve K influential concepts for an input x

$\{q\}_{1:K}$

$$d(x, Q) = \frac{x \cdot q}{\|x\| \|q\|}$$

$q \in Q$

Classification

$$q_K = \sum_{k=1}^{K} \underline{w_k} q_k$$

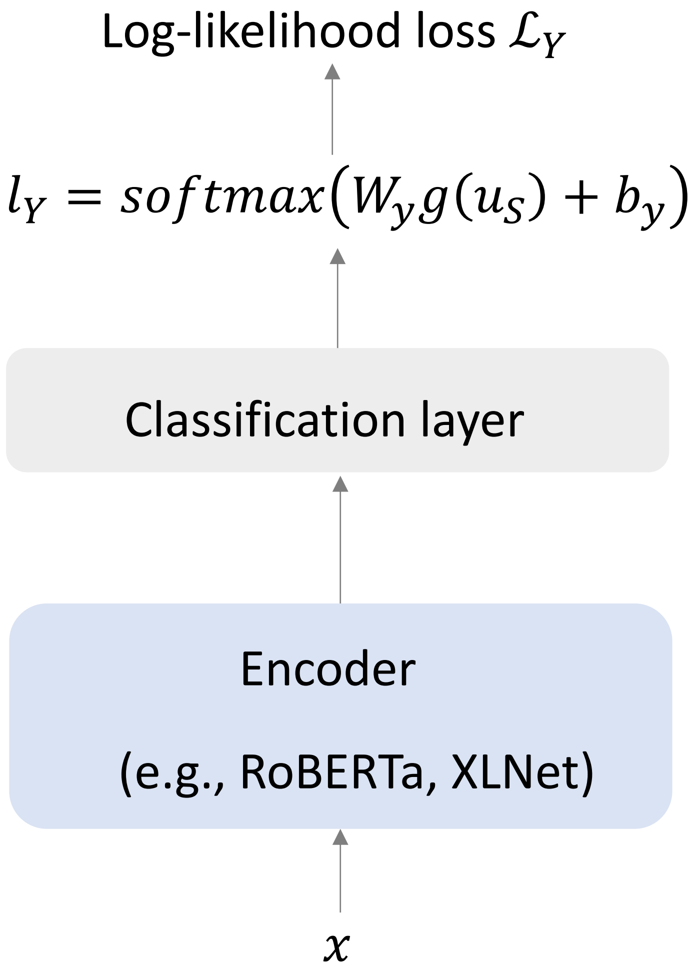
$$l_G = softmax(\underline{W_u} g(q_K) + \underline{b_u})$$

GIL parameters

Question?

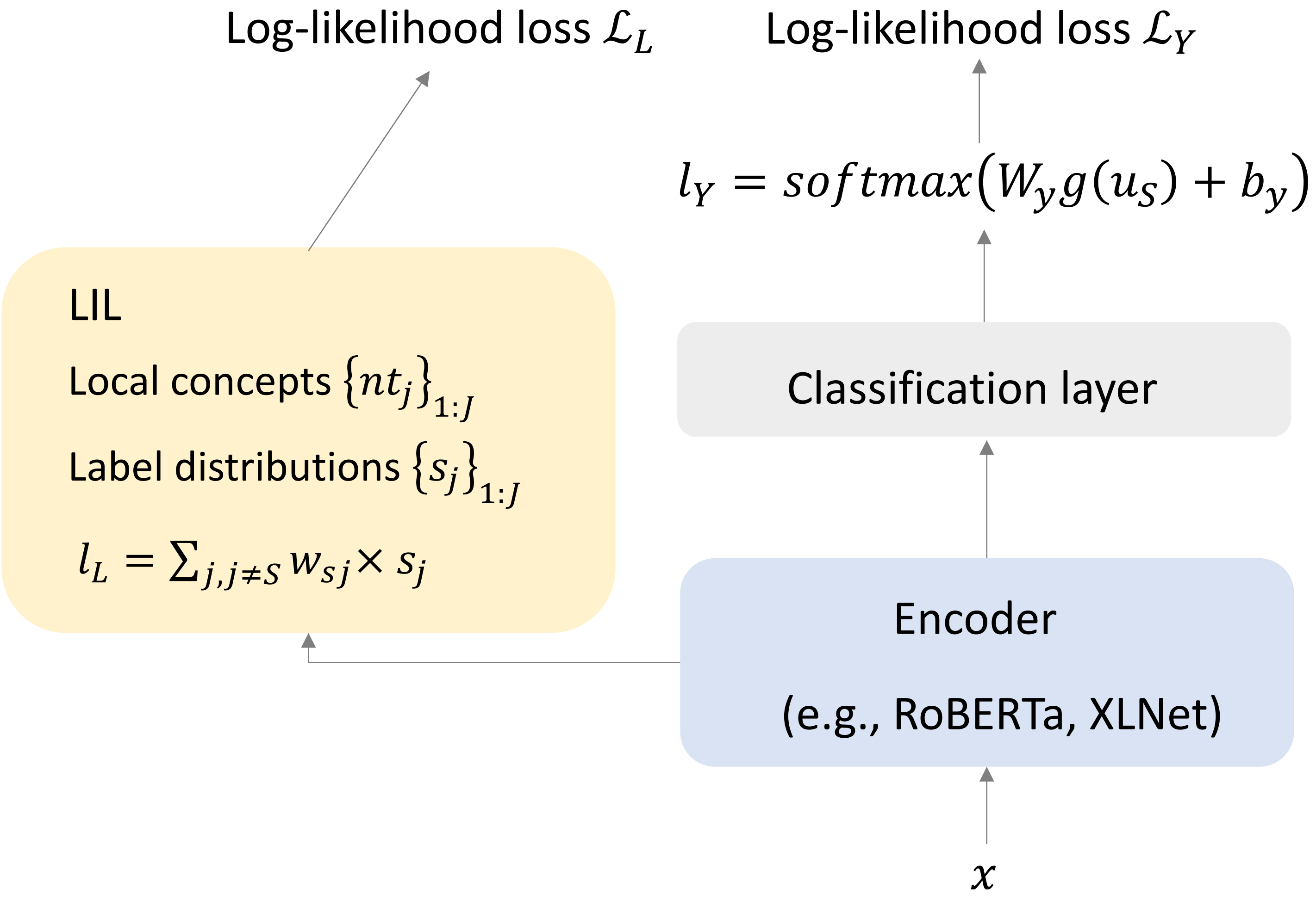
SELFEXPLAIN

Training



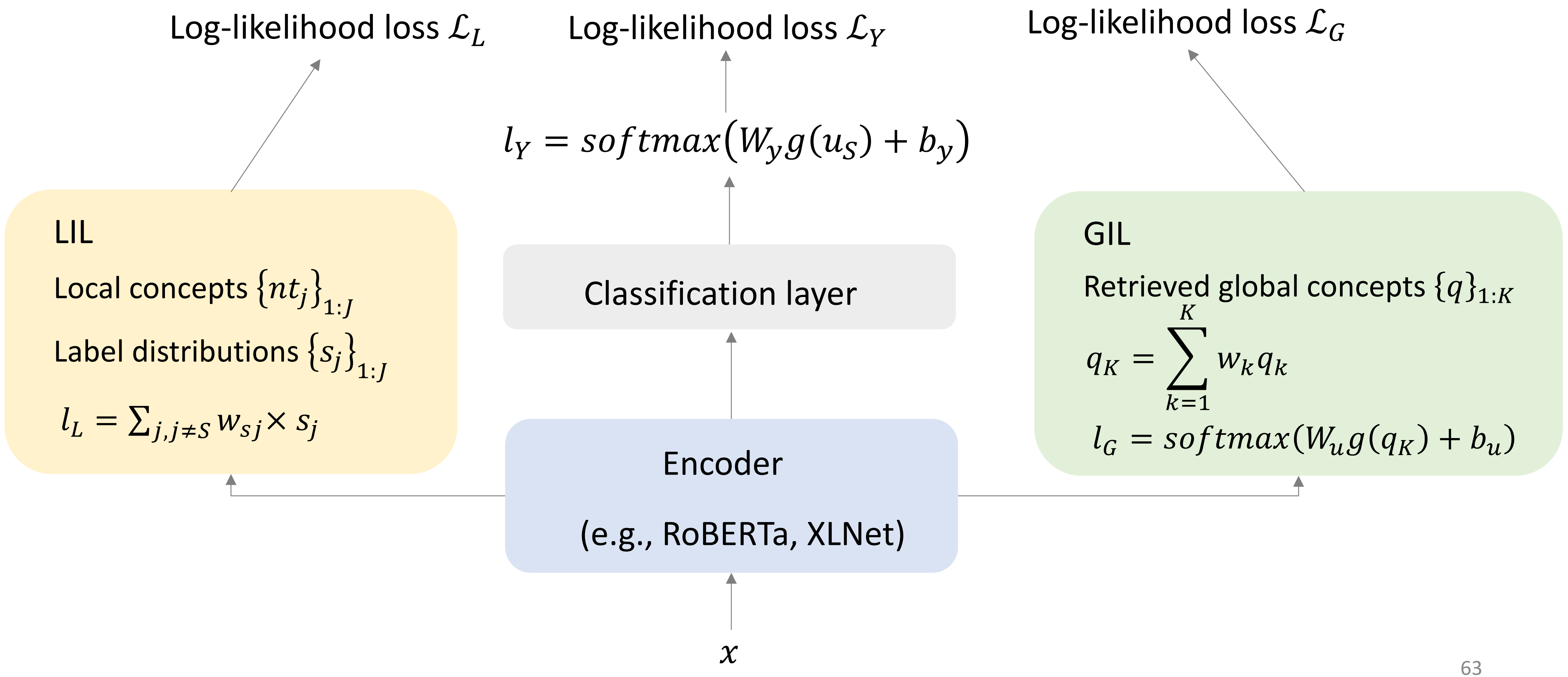
SELFEXPLAIN

Training



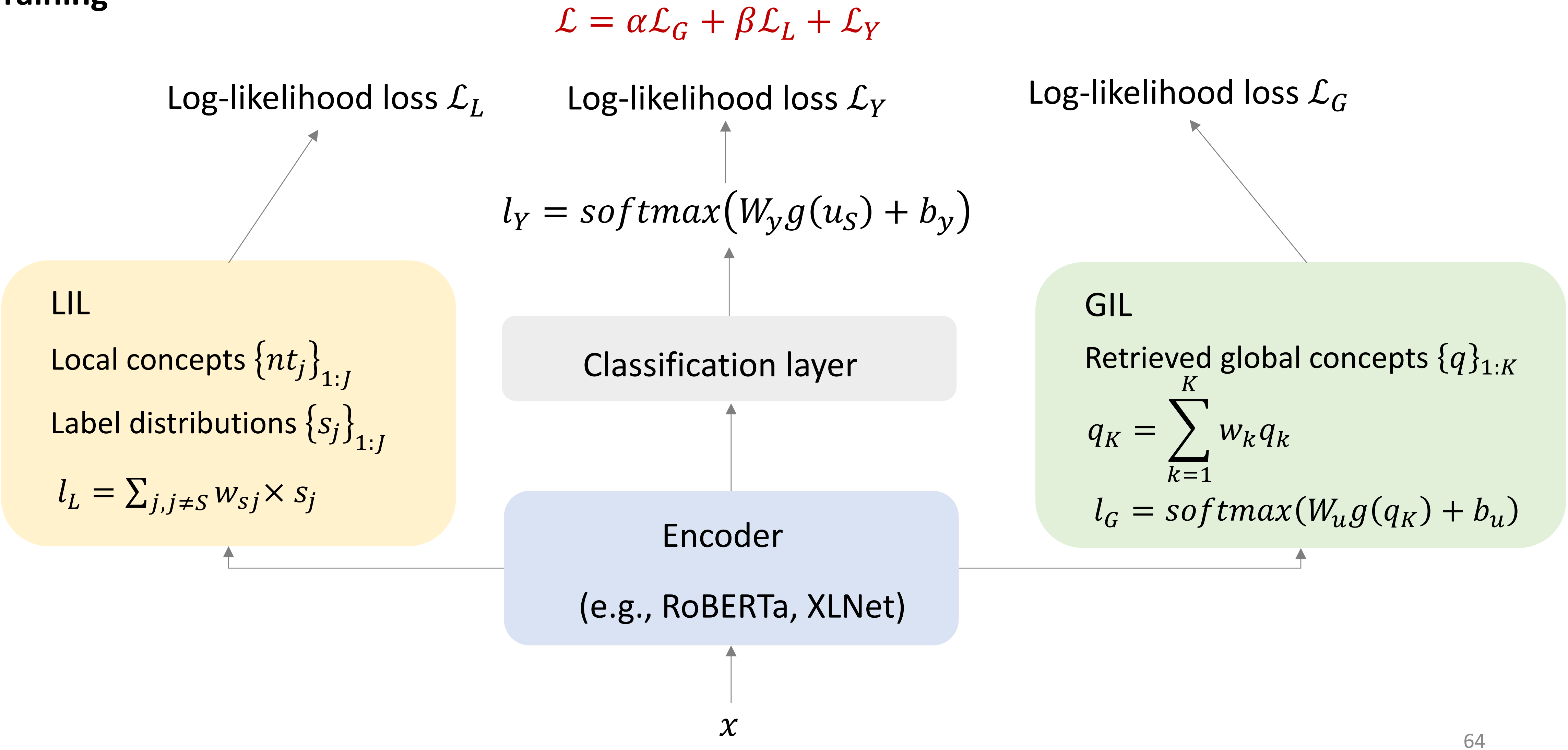
SELFEXPLAIN

Training



SELFEXPLAIN

Training



SELFEXPLAIN

Training

$$\mathcal{L} = \alpha\mathcal{L}_G + \beta\mathcal{L}_L + \mathcal{L}_Y$$

Interpretation: local relevant concepts and global influential concepts

Log-likelihood loss $\mathcal{L}_L$

Log-likelihood loss $\mathcal{L}_Y$

Log-likelihood loss $\mathcal{L}_G$

$$l_Y = softmax\left(W_y g(u_S) + b_y\right)$$

LIL

Local concepts $\{nt_j\}_{1:J}$

Label distributions $\{s_j\}_{1:J}$

$$l_L = \sum_{j, j \neq S} w_{sj} \times s_j$$

Classification layer

GIL

Retrieved global concepts $\{q\}_{1:K}$

$$q_K = \sum_{k=1}^{K} w_k q_k$$

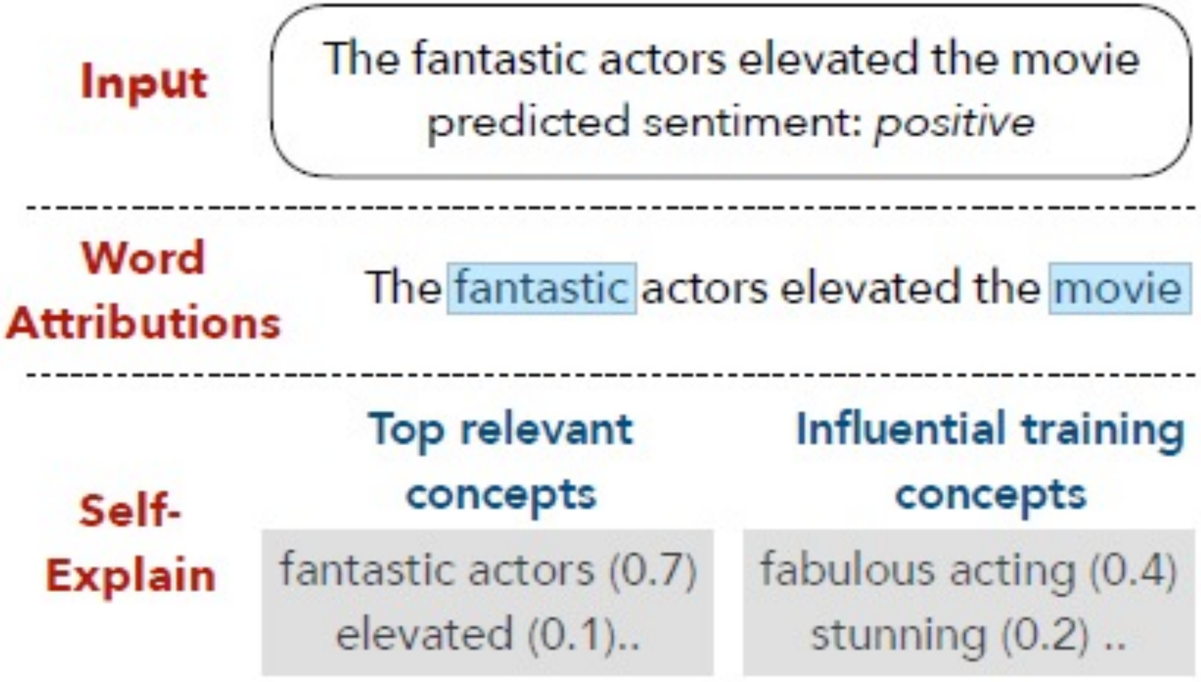
$$l_G = softmax(W_u g(q_K) + b_u)$$

Encoder

(e.g., RoBERTa, XLNet)

x

SELFEXPLAIN



Question?

Experiments

Classification performance

Comparable performance to base models across 5 text classification tasks

Model	SST-2	SST-5	TREC-6	TREC-50	SUBJ
XLNet	93.4	53.8	**96.6**	82.8	96.2
SELFEXPLAIN-XLNet (K=5)	**94.6**	**55.2**	96.4	**83.0**	**96.4**
SELFEXPLAIN-XLNet (K=10)	94.4	55.2	96.4	82.8	96.4
RoBERTa	94.8	53.5	97.0	89.0	96.2
SELFEXPLAIN-RoBERTa (K=5)	**95.1**	**54.3**	**97.6**	**89.4**	**96.3**
SELFEXPLAIN-RoBERTa (K=10)	95.1	54.1	97.6	89.2	96.3

Experiments

Explanation evaluation (local relevant concepts, global influential concepts)

- Sufficiency – Do explanations sufficiently reflect the model predictions?
- Plausibility – Do explanations appear plausible and understandable to humans?
- Trustability – Do explanations improve human trust in model predictions?

Experiments

Explanation evaluation (local relevant concepts, global influential concepts)

Sufficiency – Do explanations sufficiently reflect the model predictions?

explanations → Model → Label ?

An explanation that achieves high accuracy using this classifier is indicative of its ability to recover the original model prediction

Experiments

Explanation evaluation (local relevant concepts, global influential concepts)

Sufficiency – Do explanations sufficiently reflect the model predictions?

explanations → Model → Label ?

An explanation that achieves high accuracy using this classifier is indicative of its ability to recover the original model prediction

Model	Explanation	Accuracy
Full input text	-	0.90
Lei et al. (2016)	contiguous	0.71
	top-K tokens	0.74
Bastings et al. (2019)	contiguous	0.60
	top-K tokens	0.59
Li et al. (2016)	contiguous	0.70
	top-K tokens	0.68
[CLS] Attn	contiguous	0.81
	top-K tokens	0.81
SELFEXPLAIN-LIL	top-K concepts	**0.84**
SELFEXPLAIN-GIL	top-K concepts	**0.93**

Baselines: attention/gradient-based explanations

✓ Both LIL and GIL explanations show high predictive performance

✓ GIL explanations outperform full-text performance

Experiments

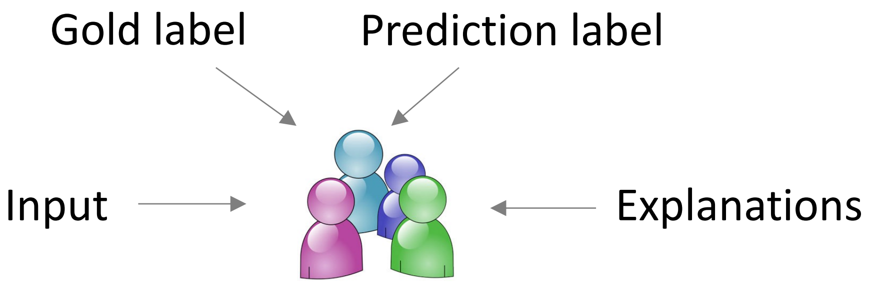
Explanation evaluation (local relevant concepts, global influential concepts)

Plausibility – Do explanations appear plausible and understandable to humans?

Trustability – Do explanations improve human trust in model predictions?

Adequate justification

Asking human judges: “Does the explanation adequately justifies the model prediction?”



Experiments

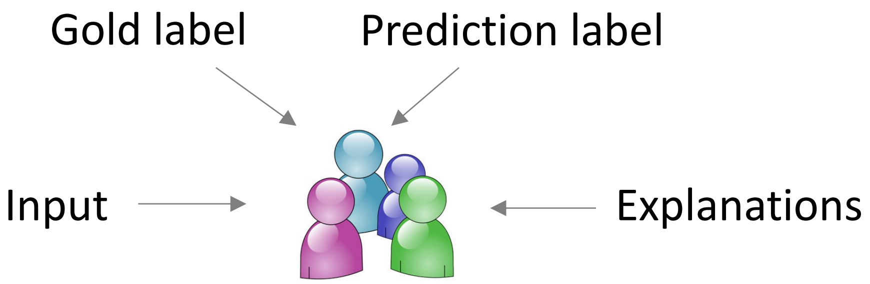
Explanation evaluation (local relevant concepts, global influential concepts)

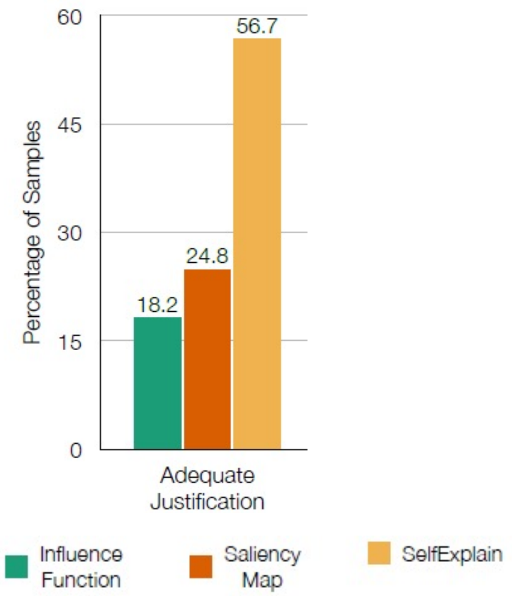
Plausibility – Do explanations appear plausible and understandable to humans?

Trustability – Do explanations improve human trust in model predictions?

Adequate justification

Asking human judges: “Does the explanation adequately justifies the model prediction?”





Experiments

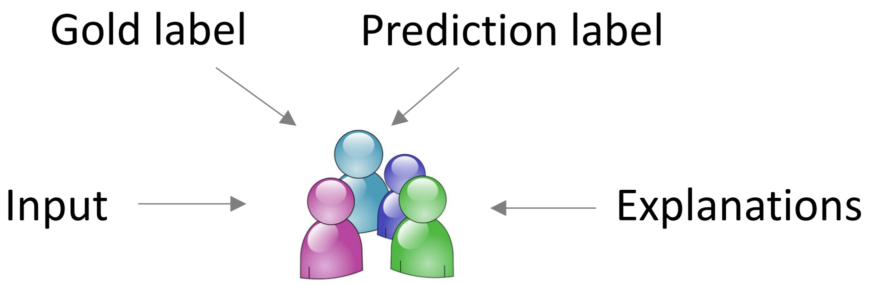
Explanation evaluation (local relevant concepts, global influential concepts)

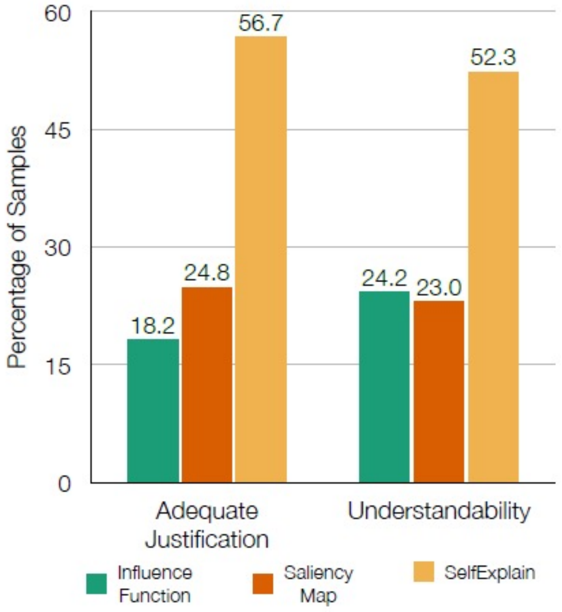
Plausibility – Do explanations appear plausible and understandable to humans?

Trustability – Do explanations improve human trust in model predictions?

Understandability

Asking human judges to select the explanations that they perceived to be more understandable





Experiments

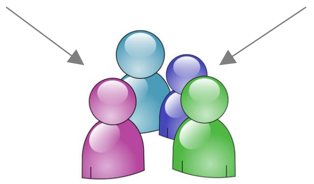
Explanation evaluation (local relevant concepts, global influential concepts)

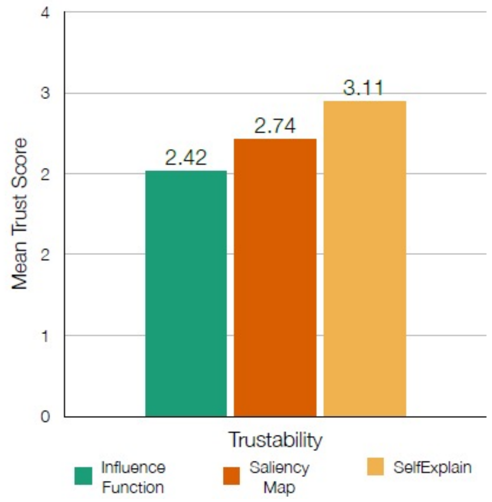
Plausibility – Do explanations appear plausible and understandable to humans?
Trustability – Do explanations improve human trust in model predictions?

Trustability

Mean trust score: asking human judges to rate on a scale of 1–5 based on how much trust each of the model explanations instill

Prediction label Explanations





Analysis

Does SELFEXPLAIN's explanation help predict model behavior?

Asking human judges to predict the model decision with and without the presence of model explanations

- ✓ When users were presented with the explanation, their ability to predict model decision improved by an average of 22%

Analysis

Global interpretations seem more reasonable

Sample	P_C	Top relevant phrases from LIL	Top influential concepts from GIL
the iditarod lasts for days - this just felt like it did .	neg	for days	exploitation piece, heart attack
corny, schmaltzy and predictable, but still manages to be kind of heart warming, nonetheless.	pos	corny, schmaltzy, of heart	successfully blended satire, spell binding fun
suffers from the lack of a compelling or comprehensible narrative .	neg	comprehensible, the lack of	empty theatres, tumble weed
the structure the film takes may find matt damon and ben affleck once again looking for residuals as this officially completes a good will hunting trilogy that was never planned .	pos	the structure of the film	bravo, meaning and consolation

Analysis

Global interpretations are more stable to input perturbations

Input	Top LIL interpretations	Top GIL interpretations
it 's a very charming and often affecting journey	often affecting, very charming	scenes of cinematic perfection that steal your heart away, submerged, that extravagantly
it ' s a charming and often affecting journey of people	of people, charming and often affecting	scenes of cinematic perfection that steal your heart away, submerged, that extravagantly

Question?

Reference

- Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." *Advances in neural information processing systems* 31 (2018).
- Rajagopal, Dheeraj, et al. "Selfexplain: A self-explaining architecture for neural text classifiers." *arXiv preprint arXiv:2103.12279* (2021).