Learning to Explain: An Information-Theoretic Perspective on Model Interpretation

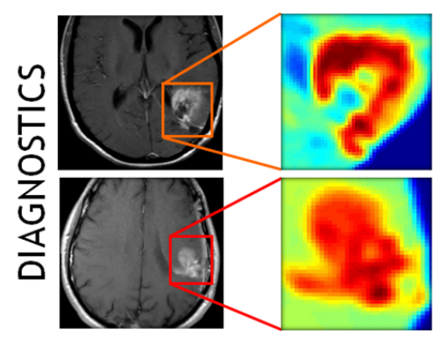
Jianbo Chen*[2] Le Song[‡,§] Martin J. Wainwright* Michael I. Jordan*

University of California, Berkeley*, Georgia Institute of Technology[‡], Ant Financial[§]

Motivation

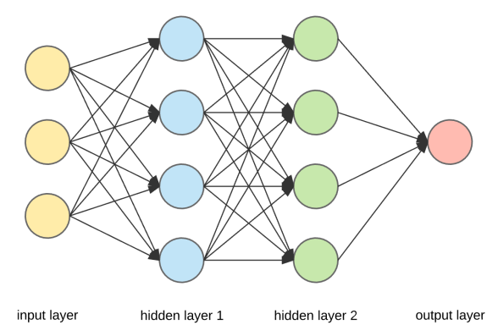
Application of machine learning:

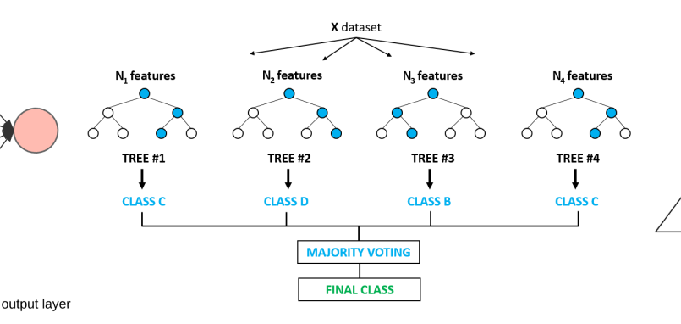
- Medicine
- Financial markets
- Criminal justice

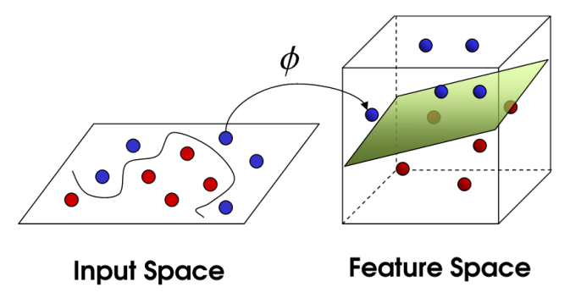


Complex models:

- Deep neural networks
- Random forests
- Kernel methods



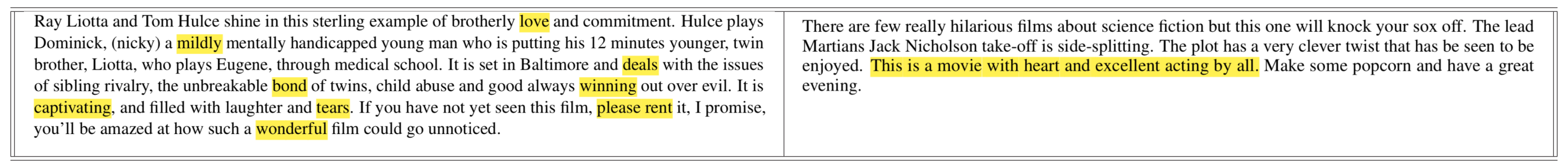




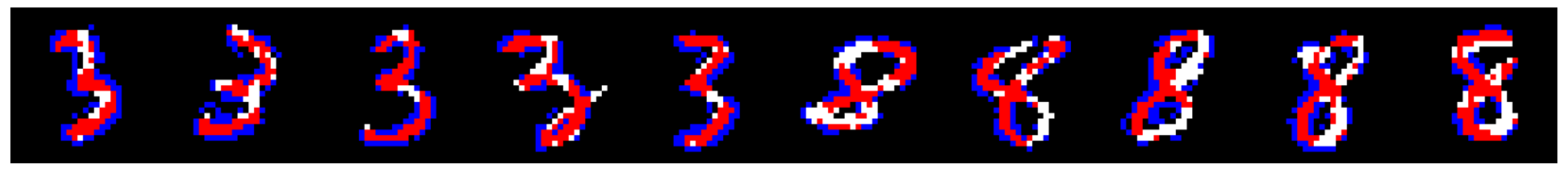
Problem

Instancewise Feature Selection:

1. Given a machine learning model, one asks for the importance score of each feature on the prediction of a given instance.
2. Feature importance is allowed to vary across instances.



Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.



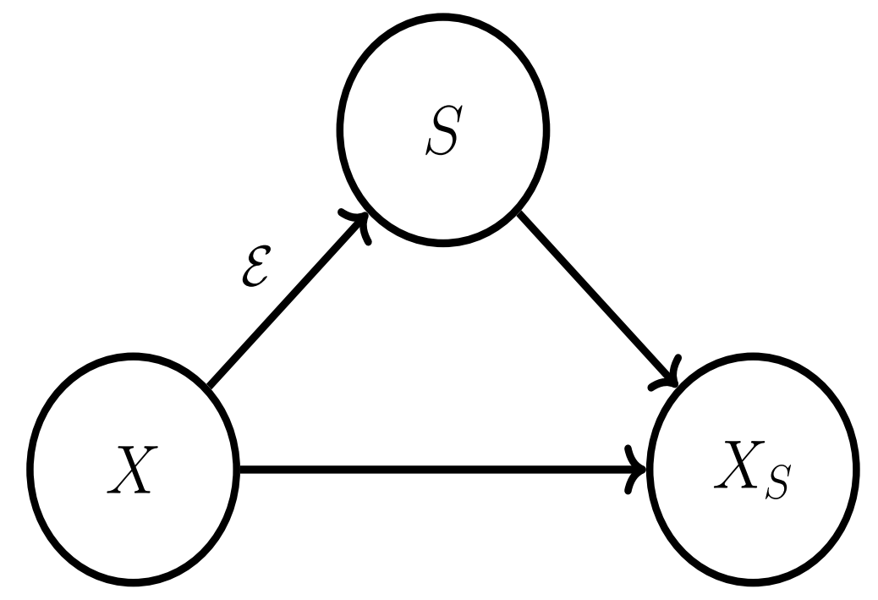
Existing Work and Properties

Training: Require training in advance.
Efficiency: Scalable to large data sets.
Additive: Approximated by an additive model locally.
Model-agnostic: Generic to black-box models.

	Training	Efficiency	Additive	Model-agnostic
Parzen (Baehrens et al., 2010)	Yes	High	Yes	Yes
Salient Map (Simonyan et al., 2013)	No	High	Yes	No
LRP (Bach et al., 2015)	No	High	Yes	No
LIME (Ribeiro et al., 2016)	No	Low	Yes	Yes
DeepLIFT (Shrikumar et al., 2017)	No	High	Yes	No
Kernel SHAP (Lundberg & Lee, 2017)	No	Low	Yes	Yes
IG (Sundararajan et al., 2017)	No	Medium	Yes	No
L2X	**Yes**	**High**	**No**	**Yes**

Notations

Input $x \in \mathbb{R}^d$.
Model $(Y \mid x) \sim \mathbb{P}_m(\cdot \mid x)$.
S: A feature subset of size k.
$\wp_k$: All subsets of size k.
Explainer $\mathcal{E}$: $\mathbb{P}(S \mid x)$.
X_S: The sub-vector of chosen features.



Framework

Objective: Maximize the mutual information between selected features and theresponse variable, over the explainer $\mathcal{E}$:

$$\max_{\mathcal{E}} I(X_S; Y) \quad \text{subject to} \qquad S \sim \mathcal{E}(X). \tag{1}$$

An information-theoretic interpretation: Define

$$\mathcal{E}^*(x) := \arg\min_S \quad \underbrace{\mathbb{E}_m\left[\log \frac{1}{\mathbb{P}_m(Y \mid x_S)} \,\middle|\, x\right]}_{\text{Expected length of encoded message for the target model using } \mathbb{P}_m(Y|x_S).}.$$

Then $\mathcal{E}^*$ is a global optimum of Problem (1). Conversely, any global optimum of Problem (1) degenerates to $\mathcal{E}^*$ almost surely over $\mathbb{P}_X$.

Intractability of the objective:

$$\begin{aligned} I(X_S; Y) &= \mathbb{E}\Big[\log \frac{\mathbb{P}_m(X_S, Y)}{\mathbb{P}(X_S)\mathbb{P}_m(Y)}\Big] = \mathbb{E}\Big[\log \frac{\mathbb{P}_m(Y \mid X_S)}{\mathbb{P}_m(Y)}\Big] \\ &= \mathbb{E}\Big[\log \mathbb{P}_m(Y \mid X_S)\Big] + \text{ Const.} \\ &= \mathbb{E}_X \underbrace{\mathbb{E}_{S|X}\mathbb{E}_{Y|X_S}\Big[\log \mathbb{P}_m(Y \mid X_S)\Big]}_{\text{Intractable to compute directly.}} + \text{Const.} \end{aligned}$$

Approximations

A variational formulation: Introduce a variational family for approximation:

$$\mathcal{Q} := \Big\{\mathbb{Q} \mid \mathbb{Q} = \{x_S \to \mathbb{Q}_S(Y|x_S), S \in \wp_k\}\Big\}.$$

An application of Jensen's inequality yields the lower bound

$$\begin{aligned} \mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y \mid X_S)] &\geq \int \mathbb{P}_m(Y \mid X_S) \log \mathbb{Q}_S(Y \mid X_S) \\ &= \mathbb{E}_{Y|X_S}[\log \mathbb{Q}_S(Y \mid X_S)], \end{aligned}$$

where equality holds iff $\mathbb{P}_m(Y \mid X_S) \overset{d}{=} \mathbb{Q}_S(Y \mid X_S)$.

A single neural network g_α for parametrizing $\mathbb{Q}$:
Define $\mathbb{Q}_S(Y|x_S) := g_\alpha(\widetilde{x}_S, Y)$, where $\widetilde{x}_S \in \mathbb{R}^d$ is defined by

$$(\widetilde{x}_S)_i = \mathbf{1}\{i \in S\} \cdot x_i.$$

Continuous relaxation of subset sampling:
Gumbel(0, 1): $G_i = -\log(-\log u_i)$, $u_i \sim \text{Uniform}(0, 1)$.
Concrete($\log p_1, \ldots, \log p_d$): A random vector $C \in \mathbb{R}^d$, with

$$C_i = \frac{\exp\{(\log p_i + G_i)/\tau\}}{\sum_{j=1}^d \exp\{(\log p_j + G_j)/\tau\}}.$$

Approximate k out of d subset sampling:

$$C^j \sim \text{Concrete}(w_\theta(X)) \text{ i.i.d. for } j = 1, 2, \ldots, k,$$

$$V(\theta, \zeta) = (V_1, V_2, \ldots, V_d), \qquad V_i = \max_j C_i^j,$$

$$\widetilde{X}_S \approx V(\theta, \zeta) \odot X.$$

(τ: temperature, θ: parameters of explainer, ζ: auxiliary random variables, $\odot$: elementwise product)

Final Objective

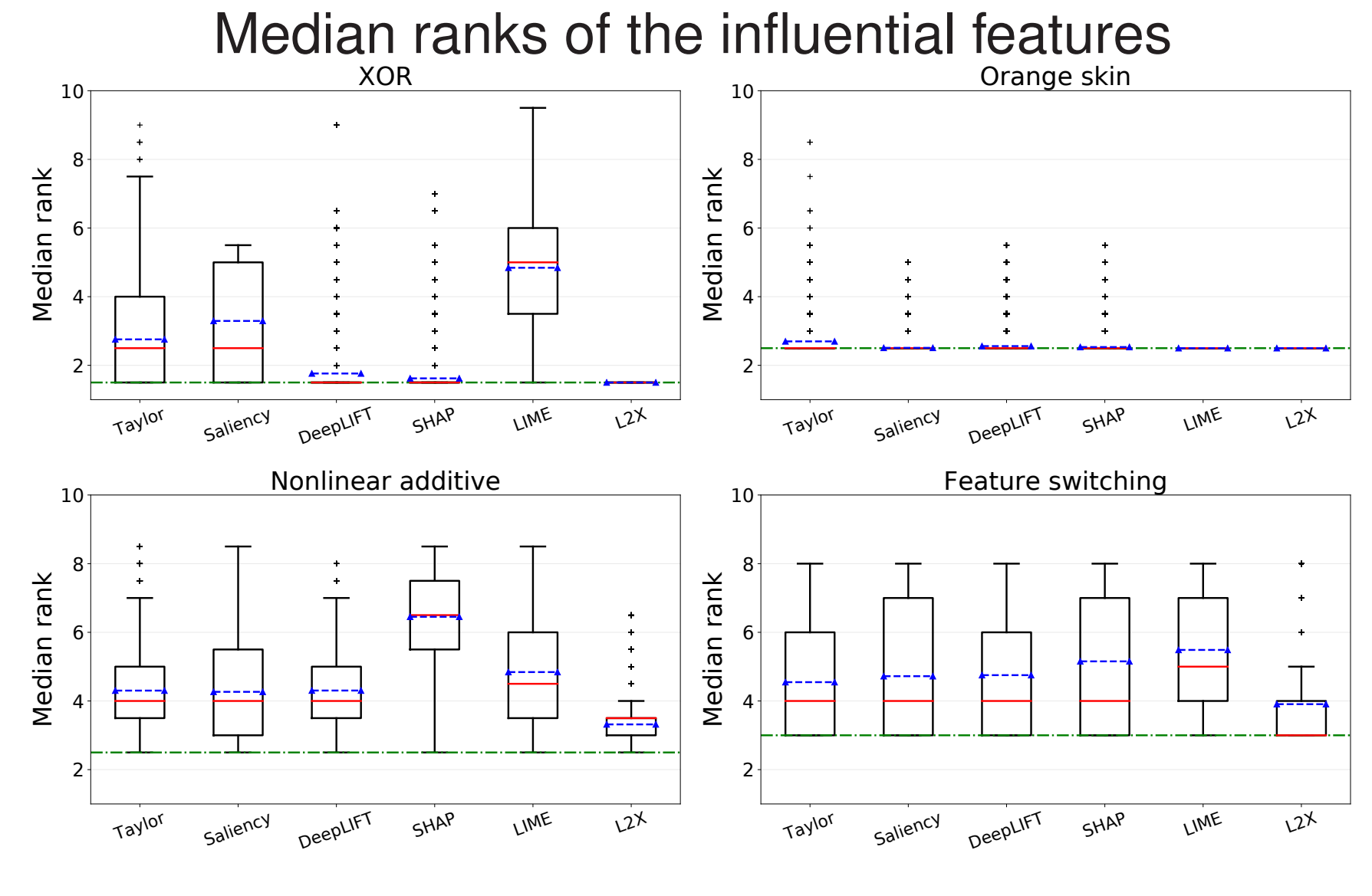
Objective: Containing parameters of both explainer and variational dist. θ, α.

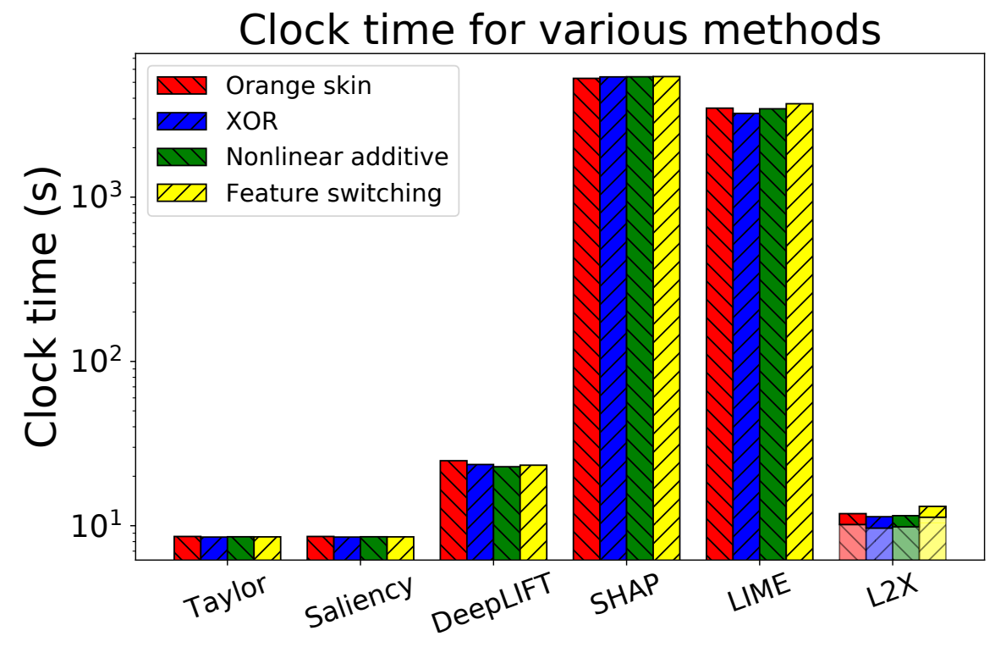
$$\max_{\theta,\alpha} \mathbb{E}_{X,Y,\zeta}\Big[\log g_\alpha(V(\theta, \zeta) \odot X, Y)\Big].$$

Optimization: Stochastic gradient methods such as Adam and RMSProp.

Synthetic Experiments

Four data sets: Orange skin, XOR, Nonlinear additive model, Switch.
Comparing methods: Saliency Map, DeepLIFT, KernelSHAP, LIME.
Evaluation: Median rank of the influential features, time complexity.





(The training time of L2X is shown in translucent bars.)

(Dotted green: optimal ranks; red: median; dotted blue: mean.)

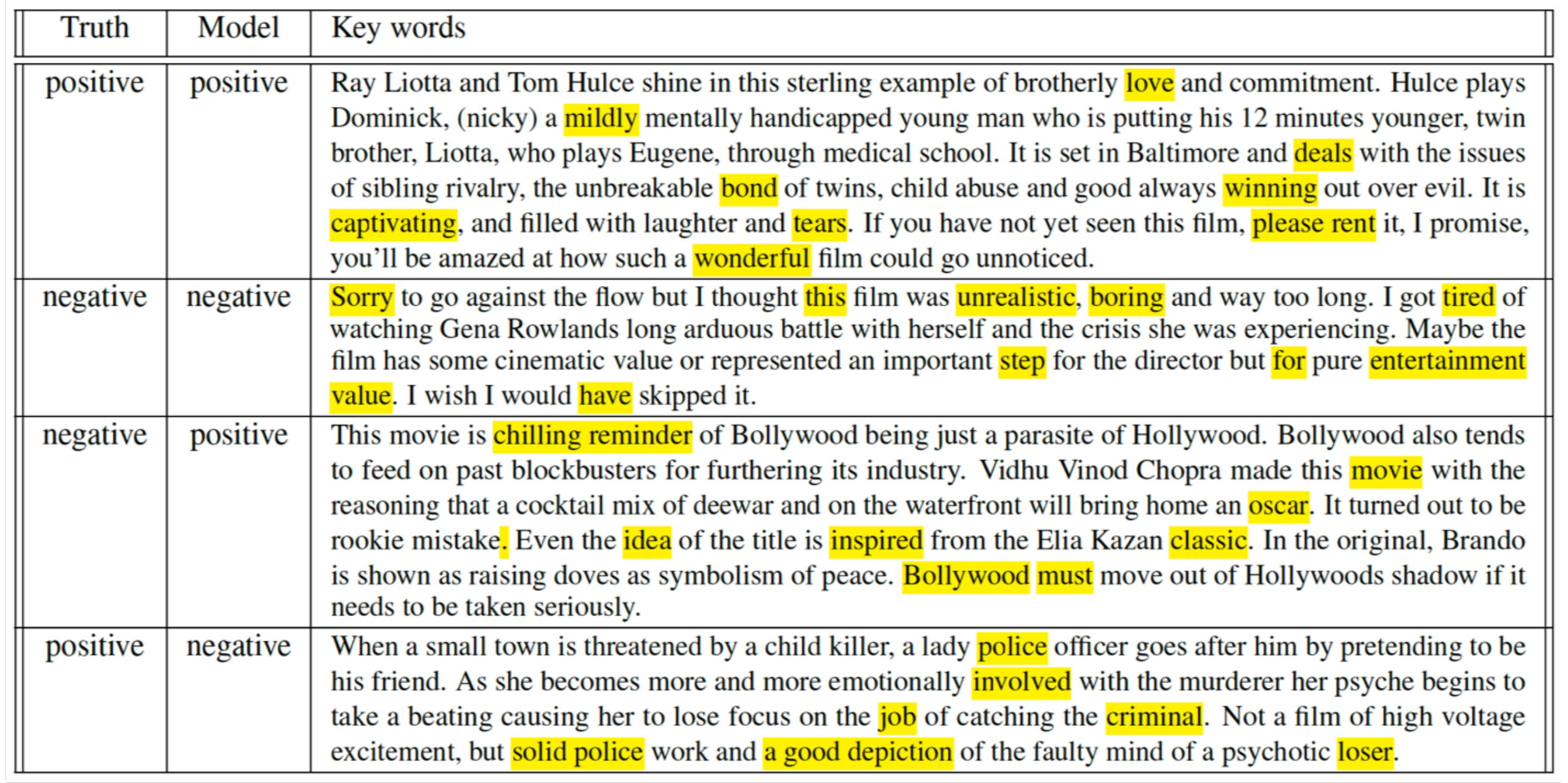
Real-world Experiments

Data sets and models: IMDB movie review with word-based CNN and Hierarchical LSTM respectively, MNIST with CNN.
Evaluation: Post-hoc accuracy, human accuracy.

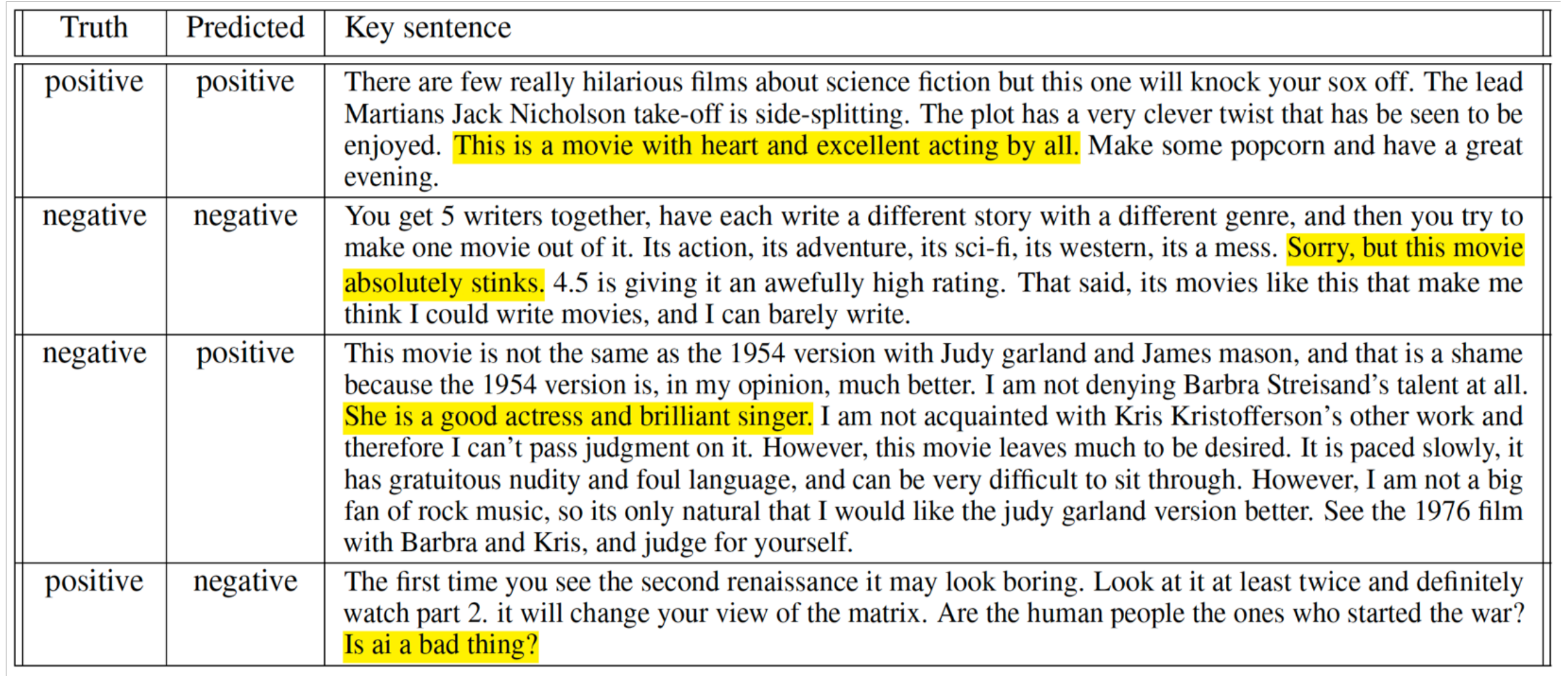
	IMDB-Word	IMDB-Sent	MNIST
Post-hoc accuracy (PA)	0.90.8	0.849	0.958
Human accuracy (HA)	0.844	0.774	NA

HA on words: 84.4% > HA on original: 83.7%

Visualization:



Truth	Model	Key words
positive	positive	Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.
negative	negative	Sorry to go against the flow but I thought this film was unrealistic, boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the crisis she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value. I wish I would have skipped it.
negative	positive	This movie is chilling reminder of Bollywood being just a parasite of Hollywood. Bollywood also tends to feed on past blockbusters for furthering its industry. Vidhu Vinod Chopra made this movie with the reasoning that a cocktail mix of deewar and on the waterfront will bring home an oscar. It turned out to be rookie mistake. Even the idea of the title is inspired from the Elia Kazan classic. In the original, Brando is shown as raising doves as symbolism of peace. Bollywood must move out of Hollywoods shadow if it needs to be taken seriously.
positive	negative	When a small town is threatened by a child killer, a lady police officer goes after him by pretending to be his friend. As she becomes more and more emotionally involved with the murderer her psyche begins to take a beating causing her to lose focus on the job of catching the criminal. Not a film of high voltage excitement, but solid police work and a good depiction of the faulty mind of a psychotic loser.



Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.
negative	negative	You get 5 writers together, have each write a different story with a different genre, and then you try to make one movie out of it. Its action, its adventure, its sci-fi, its western, its a mess. Sorry, but this movie absolutely stinks. 4.5 is giving it an awefully high rating. That said, its movies like this that make me think I could write movies, and I can barely write.
negative	positive	This movie is not the same as the 1954 version with Judy garland and James mason, and that is a shame because the 1954 version is, in my opinion, much better. I am not denying Barbra Streisand's talent at all. She is a good actress and brilliant singer. I am not acquainted with Kris Kristofferson's other work and therefore I can't pass judgment on it. However, this movie leaves much to be desired. It is paced slowly, it has gratuitous nudity and foul language, and can be very difficult to sit through. However, I am not a big fan of rock music, so its only natural that I would like the judy garland version better. See the 1976 film with Barbra and Kris, and judge for yourself.
positive	negative	The first time you see the second renaissance it may look boring. Look at it at least twice and definitely watch part 2. it will change your view of the matrix. Are the human people the ones who started the war? Is ai a bad thing?