Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

DNN中的理论可解释性

博士研究生 段学明

2024年01月07日

问题回溯

- 相关内容
  - 2021.07.11 慕星星《图神经网络可解释方法》
  - 2020.11.22 董勃《法律文本可解释性研究》
  - 2020.10.25 慕星星《机器学习常用的可解释方法》

内容提要

- 预期收获
- 内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - 博弈交互
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 从理论角度认识DNN的可解释性
  - 理解构建DNN可解释性体系的基础理论
  - 理解理论可解释的必要性
  - 了解理论可解释性的未来工作

内涵解析

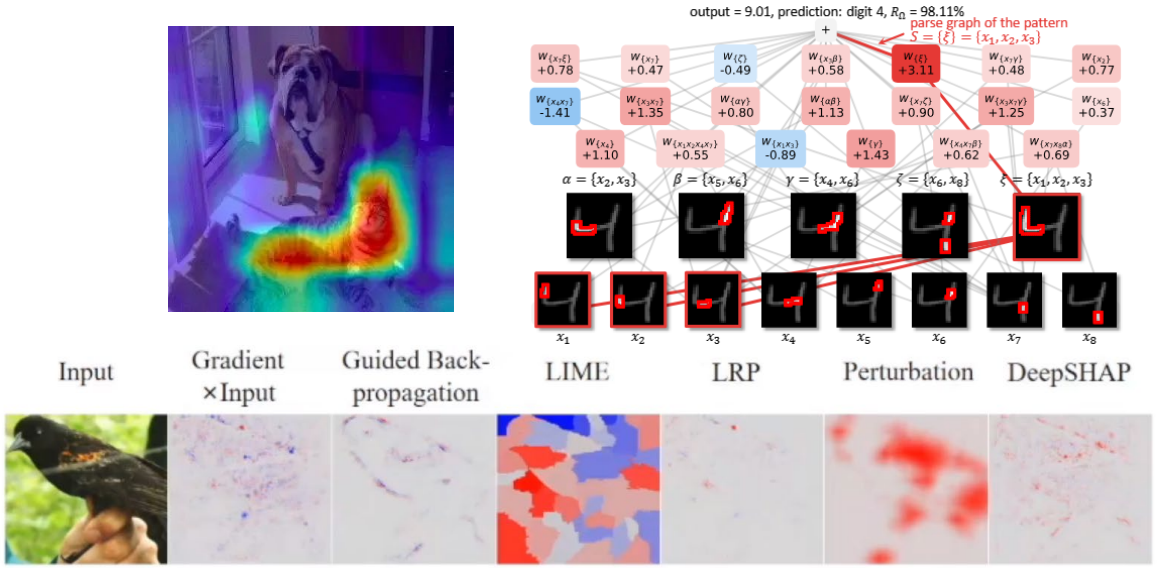
- 研究目标
  - 将可解释性从工程化解释转为对机理的探索
  - 量化DNN所建模的“知识”
  - 解释DNN的表征能力
  - 根本目标：形成可证明、可测量、可验证的第一性原理
- 理论（数学）可解释性
  - 找到上述的第一性原理

研究背景与意义

- 背景
  - 自DNN的出现起，人们不断尝试去解释
  - 随着DNN模型在重要领域的应用，人们急需得到DNN内部的运行机理
  - DNN模型变得越来越大、越来越复杂、越来越黑盒
- 必要性
  - 最基本的应用：能否解释清楚并不重要
  - 重要领域的重要问题（金融、军事、医疗）：需要可靠性
  - 目前依靠直觉
    - 如何证明某个特定领域中，一个结构比另一个结构好：并非做实验大量测试
    - 如何证明结构和知识表达存在关系：最好的结构不一定是最好的表达
    - 提高模型性能的算法的设计初衷：有效但初衷是直觉

研究历史与现状

- 语义层面解释
  - 建模了哪些概念并用于预测
  - 量化并提升可靠性
  - 端到端自动学习可被解释的特征
  - 语义层面的交流
  - 如何评价解释性
- 数学层面解释
  - 建模并解释表示能力：泛化性、鲁棒性等
  - 结构与知识表达间的关系
  - 解释经典技术：蒸馏、对抗学习等
  - 利用特征去诊断调试



只能自圆其说，不能相互印证

受限于假设、无法提供语义解释、不统一

- 内涵解释
  - 合作博弈中，用于公平地定量评估参与者贡献度的指标
  - 例子：一项工作
    - $a1$单独工作收益为8，$a2$收益为4，$a3$收益为2
    - $a1 + a2$： 二人合作的收益为20
    - $a1 + a3$： 二人合作的收益为8
    - $a2 + a3$： 二人合作的收益为12
    - 三人一起： 合作的收益为30
    - 如何分配?

合作中的公平分配

- 内涵解释
  - 价值$v(\{1\})=8, v(\{2\})=4, v(\{3\})=2$

    $v(\{1,2\})=20, v(\{1,3\})=8, v(\{2,3\})=12$

    $v(\{1,2,3\})=30$
  - 边际贡献$\Delta i(S)=v(S\cup\{i\})-v(S)$
  - $\varphi_1(v)=\frac{1}{6}(8+8+16+18+6+18)$

    $=\frac{74}{6}$
  - 回报$\omega_1=\omega_2=41.1\%, \omega_3=17.8\%$

P	Order	Delta 1	Delta 2	Delta 3
1/6	123	8	$v(12)-v(1)=12$	$v(123)-v(12)=10$
1/6	132	8	$v(123)-v(13)=22$	$v(13)-v(1)=0$
1/6	213	$v(12)-v(2)=16$	4	$v(123)-v(12)=10$
1/6	231	$v(123)-v(23)=18$	4	$v(23)-v(2)=8$
1/6	312	$v(13)-v(3)=6$	$v(123)-v(13)=22$	2
1/6	321	$v(123)-v(23)=18$	$v(23)-v(3)=10$	2

- 数学解释
  - 设所有参与者的集合N，N的所有潜在子集$2^N \stackrel{\text{def}}{=} \{S|S \subseteq N\}$。对于每个子集$S$，一场博弈$v$: $2^N \to \mathbb{R}$表示由S获得的标量奖励。
  - 参与者i的Shapley值$\varphi(i)$表示i对博弈v的贡献：

$$\varphi(i|N) = \sum_{S \subseteq N \setminus \{i\}} \frac{(|N| - |S| - 1)!\,|S|!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

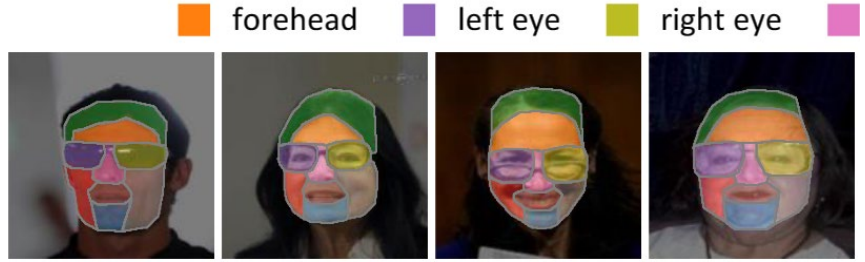
Shapley值是一种无偏方法

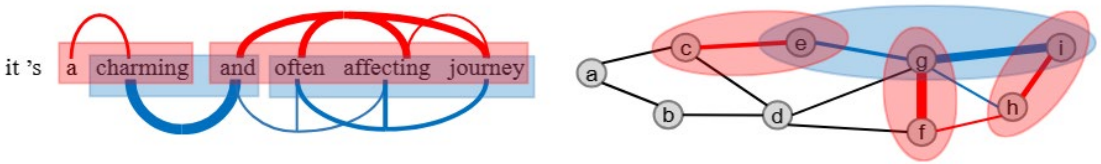
- 数学性质
  - 可加性：$\forall S \subseteq N, \varphi_u(i|N) = \varphi_v(i|N) + \varphi_w(i|N)$
  - 伪玩家：$\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\}),\ \varphi(i|N) = v(\{i\}) - v(\emptyset)$
    - i在N中与其他参与者没有互动
  - 对称性：$\text{If } \forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S \cup \{j\}), \text{then } v(i|N) = v(j|N)$
    - 可交换
  - 有效性：$\sum_{i \in N} \varphi(i|N) = v(N) - v(\emptyset)$
    - 所有参与者的贡献和=总合作的价值

有且仅有一个φ同时满足4个性质，即Shapley Value

- 现象：$1+1>2$
  - 输入单元组合的重要性高，形成了交互，且有正负

it ' s a remarkably solid and subtly satirical tour de force .
this is a good script , good dialogue , funny even for adults
dull, lifeless, and amateurishly assembled .
a warm but realistic meditation on friendship , family and affection .
no telegraphing is too obvious or simplistic for this movie .
a tender , witty , captivating film about friendship , love , memory , trust and
a dumb movie with dumb characters doing dumb things and you have to be





  - 理论上保证解释的严谨性

$$v(N) = I(S_1) + I(S_2) + I(S_3) + I(S_4) + \dots + v(\emptyset)$$

Interpreting Multivariate Shapley Interactions in DNNs

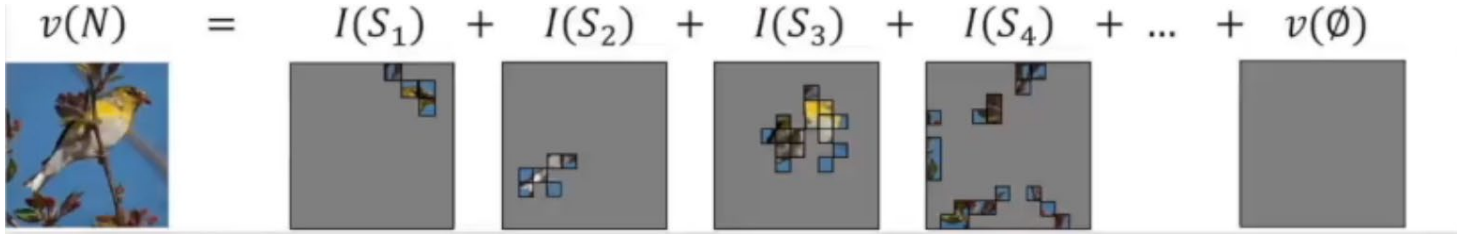
博弈交互 TIPO

T	目标	从多元相互作用的角度解释经过训练的DNN
I	输入	DNN原始输入与不同遮盖
P	处理	1.定义多变元博弈交互 2.近似计算多变元博弈交互
O	输出	量化的DNN的多个输入变量之间相互作用的重要性

P	问题	缺少对DNN建模的语义的数学建模
C	条件	Shapley Value理论
D	难点	近似多变元博弈交互的高准确性和证明
L	水平	AAAI 2021

理论构建　双变元博弈交互

- 基于博弈论定义参与者和博弈
  - $N = \{1, 2, \ldots, n\}$，　$2^N = \{S | S \subseteq N\}$，　v: $2^N \to \mathbb{R}$
  - If $N = \{a, b\}$,　then $2^N = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$
- 关注S
  - $v(S), S \subseteq N$，表示$N \backslash S$中所有输入变量都被基线值替换时DNN的输出



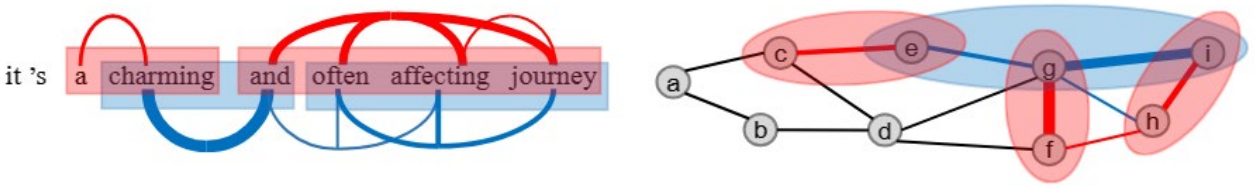
将具有标量输出的DNN视为一个博弈，将输入变量视为一组参与者

理论构建　双变元博弈交互

- v中i与j的交互
  - $i, j \rightarrow \varphi(i|N), \varphi(j|N)$
  - i、j总是一起参与博弈，或总是一起缺席：认为i、j组成了一个联盟S_{ij}
  - i、j单独参与博弈的贡献和往往≠联盟S_{ij}的贡献
  - $\varphi(S_{ij}|N)$计算：$\varphi(i|N) = \sum_{S \subseteq N \setminus \{i\}} \frac{(|N|-|S|-1)!|S|!}{|N|!}[v(S \cup \{i\}) - v(S)]$，用$S_{ij}$替换$i$
- 例子
  - $N = \{\text{He is a green hand}\}$

理论构建　双变元博弈交互

- 数学计算
  - $B(S_{ij}) = \varphi(S_{ij}|N') - [\varphi(i|N_i) + \varphi(j|N_j)] = \sum_{S \subseteq N \setminus \{i,j\}} \frac{(n-|S|-2)!|S|!}{(n-1)!} \Delta v(i,j,S)$
    - $N' = N \setminus \{i,j\} \cup S_{ij}$, $N_i = N \setminus \{j\}$, $N_j = N \setminus \{i\}$
    - $\Delta v(i,j,S) = v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup)\{j\} + v(S)$
  - $B(S_{ij}) > 0$：i,j之间有积极的博弈交互作用
  - $B(S_{ij}) < 0$：i,j之间有消极的博弈交互作用

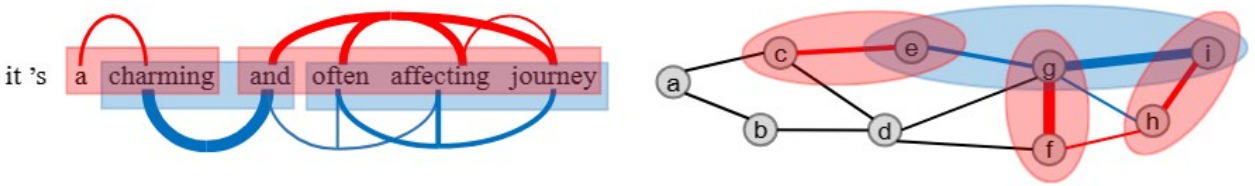


理论构建　多变元博弈交互

- 关注集合N的真子集A
  - A中成员总是一起参与博弈，或总是一起缺席：认为A组成了一个联盟$[A]$
  - $B([A]) = \varphi\left([A]\middle|N_{[A]}\right) - \sum_{i\in A}\varphi(i|N_i)$
    - $N_{[A]} = N\backslash A \cup \{[A]\},\ N_i = N\backslash A \cup \{i\}$

$B([A])$的表达全面了吗?

- 例子：$A = \{i, j, k, l\}$



- 改良指标：同时反映A中正的和负的交互
  - 划分联盟A：$\Omega = \{C_1, C_2, \dots, C_k\}$ →找到一个划分Ω，使其中的联盟尽可能多地包含正的交互→ $\Omega = \{C_1 = \{i, j\}, C_2 = \{k, l\}\}$
  - 将每个联盟视作单独的参与者，有：

$$B_{max}([A]) = \max_{\Omega} \sum_{C \in \Omega} \varphi(C|N_C) - \sum_{i \in A} \varphi(i|N_i)$$

    - $N_C = \mathrm{N} \backslash \mathrm{A} \cup \{[C]\},\ N_i = N \backslash A \cup \{[C]\}$
  - 类似地，有$B_{min}([A])$
  - 定义$T([A])$：量化A中正的和负的交互

$$T([A]) = B_{max}([A]) - B_{min}([A]) = \max_{\Omega} \sum_{C \in \Omega} \varphi(C|N_C) - \min_{\Omega} \sum_{C \in \Omega} \varphi(C|N_C)$$

理论证明 $B[A]$与$I(A)$、$B[A]$与$T([A])$

- 基本相互作用分量$I(A)$：量化A中各成员组成一个联盟时带来的额外贡献，同时移除了A中所有可能的子集所包含的贡献
- 证明：$B([A]) = F[I(A)]$、$B([A]) \to T([A])$
  - 给定v和集合N，$A \subseteq N$

$$I(A) = \sum_{T \subseteq N \setminus A} P_{\text{Shapley}}(T|N \setminus A) \sum_{L \subseteq A} (-1)^{|A|-|L|} v(L \cup T)$$

    - $P_{\text{Shapley}}(T|N \setminus A) = \frac{(|N|-|T|-|A|)!|T|!}{(|N|-|A|+1)!}$
  - 对于单个参与者i，$I(i) = \varphi(i)$

理论证明 $B[A]$与$I(A)$、$B[A]$与$T([A])$

- 证明：$B([A]) = F[I(A)]$、$B([A]) \to T([A])$
  - 对于$[A]$，$\forall |A| > 1$，$A \subseteq N$

$$I(A) = I^{N_{[A]}}([A]) - \sum_{K \subsetneq A,\, A \neq \emptyset} I^{N \backslash K}(A \backslash K)$$

$$= I^{N_{[A]}}([A]) - \sum_{A' \subsetneq A,\, A \neq \emptyset} I^{N_{A'}}(A')$$

  - $N_{[A]} = N \backslash A \cup \{[A]\}$，$N_{A'} = N \backslash A \cup A'$

理论证明 $B[A]$与$I(A)$、$B[A]$与$T([A])$

- 证明：$B([A]) = F[I(A)]$、$B([A]) \to T([A])$

$$B([A]) = \varphi_{[A]}^{N_{[A]}}([A]) - \sum_{i \epsilon A} \varphi_i^{N_i}(i)$$

$$= I^{N_{[A]}}([A]) - \sum_{i \epsilon A} I^{N_i}(i)$$

$$= I(A) + \sum_{A' \subsetneq A,\, A \neq \emptyset} I^{N_{A'}}(A') - \sum_{i \epsilon A} I^{N_i}(i)$$

$$= \sum_{A' \subseteq A,\, |A'| > 1} I^{N_{A'}}(A')$$

- $B([A]) \to 2^m - m - 1$

理论证明 $B[A]$与$I(A)$、$B[A]$与$T([A])$

- 推论： $B([A]) = \sum_{A' \subseteq A, |A'|>1} I(A')$
  - 交互$B([A])$中共有$2^m - m - 1$个不同的基本交互成分， $m = |A|$
  - 交互成分有正有负，量化所有交互成分的简单想法是绝对值相加

$$B'([A]) = \sum_{A' \subseteq A, |A'|>1} |I(A')| = B^+ - B^-$$

  - $|A'| > 1$下，$B^+ = \sum_{A' \subseteq A, I(A')>0} I(A')$， B^-同理
  - 结论： $B_{max}([A])$被认为主要编码A中大多数正交互，与B^+高度相关

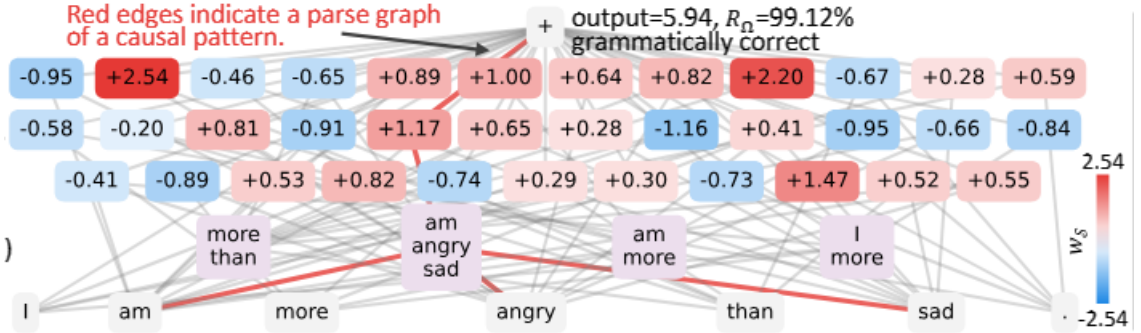
$I(A)$为所有可能的$A' \subseteq A$计算了2^m个不同的$I(A')$，$B([A])$表示所有m个输入变量之间整体交互显著性的单一度量

数学计算 采样近似

- 考虑NLP任务
  - 输入句子有单词$N=\{1,2,3,\dots,n\}$，每个单词视为一个参与者，DNN视为博弈，对某个单词子集S，将所有不在集合S中的单词遮挡，此时DNN输出即为$v(S)$
  - 某个单词之间有强烈的博弈交互，意味着DNN将其作为一个整体进行建模，将这样的联盟视为一个原型特征
  - 多元交互可以分析DNN中的原型特征
  - 非连续词通常比连续词包含更少的交互，因此要求联盟由连续单词组成

上下文的联系还存在吗?

- 计算问题
  - Ω划分情况指数级，计算成本高



数学计算 采样近似

- 近似max_{Ω}
  - 概率集合$\mathrm{p}=\{p_1,p_2,\cdots,p_{m-1}\}$表示所有可能的划分
    - $p_i(0\le p_i\le 1)$表示子集中第i个单词与第$i+1$个单词属同一联盟的概率
  - 从p中采样g，$\mathrm{g}=\{g_1,g_2,\cdots,g_{m-1}\}$，$g_i\sim\mathrm{Bernoulli}(p_i)$
  - g确定特定的Ω_g，$g_i=1$表示第i个单词与第$i+1$个单词属同一联盟
  - 有$max_{\mathrm{p}}\mathbb{E}_{\mathrm{g}\sim\mathrm{Bernoulli(p)}}\sum_{C\epsilon\Omega_{\mathrm{g}}}\varphi(C|N_C)$近似$max_{\Omega}\sum_{C\epsilon\Omega}\varphi(C|N_C)$

$$\mathcal{L}=\mathbb{E}_{g\sim\mathrm{Bernoulli}(p)}\sum_{C\epsilon\Omega_g}\varphi(C|N_C)$$

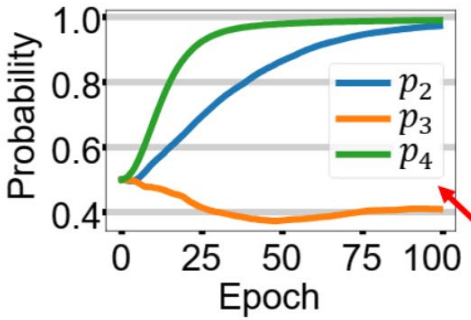
$$max_{\mathrm{p}}\mathcal{L}=max_{\mathrm{p}}\sum_{i\in A}^{\mathbb{E}}\left\{\mathbb{E}_{\mathrm{g}\sim\mathrm{Bernoulli(p)}}\frac{1}{C_i}\left[\mathbb{E}_{S\in Sam(\Omega_{\mathrm{g}}),S\ni i,|S|=r+1}[v(S)]-\mathbb{E}_{S\in Sam(\Omega_{\mathrm{g}}),S\not\ni i,|S|=r}[v(S)]\right]\right\}$$

  - 训练p：$\frac{\partial\mathcal{L}}{\partial p_i}$，近似精度随采样次数增加而增加

理想的划分对应唯一的p_i，不同遮盖→不同的采样g

实验验证　联盟划分的正确性

- 数据集
  - 构建原因：没有词间交互的基本事实注释
  - 构建：加法-乘法、与或、指数数据集
    - 加法-乘法：10000组仅由加法运算和乘法运算组成，如$y = f(x) = x1 + x2 * x3 + x4 * x5 + x6 + x7$，$x \in \{0, 1\}$
    - 多项式为黑盒，选择许多顺序输入变量构建A，如$A = \{x2, x3, x4, x5\}$
    - 最大化$\sum_{C \in \Omega} \varphi(C|N_C)$

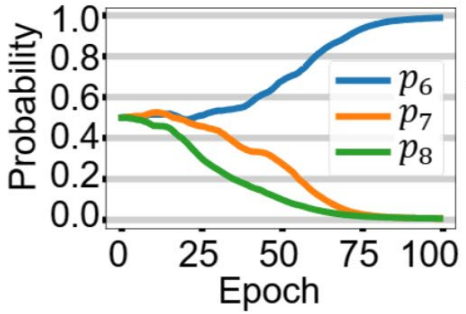


$$x_0 \times x_1 \times x_2 \times x_3 + x_4 \times x_5 + x_6 + x_7$$

p_2 p_3 p_4

$$x_0 \times x_1 \times x_2 \times x_3 + x_4 \times x_5 + x_6 + x_7$$

Theoretically, the addition operation does not cause inter-actions. Grouping additive terms does not change $T([A])$.



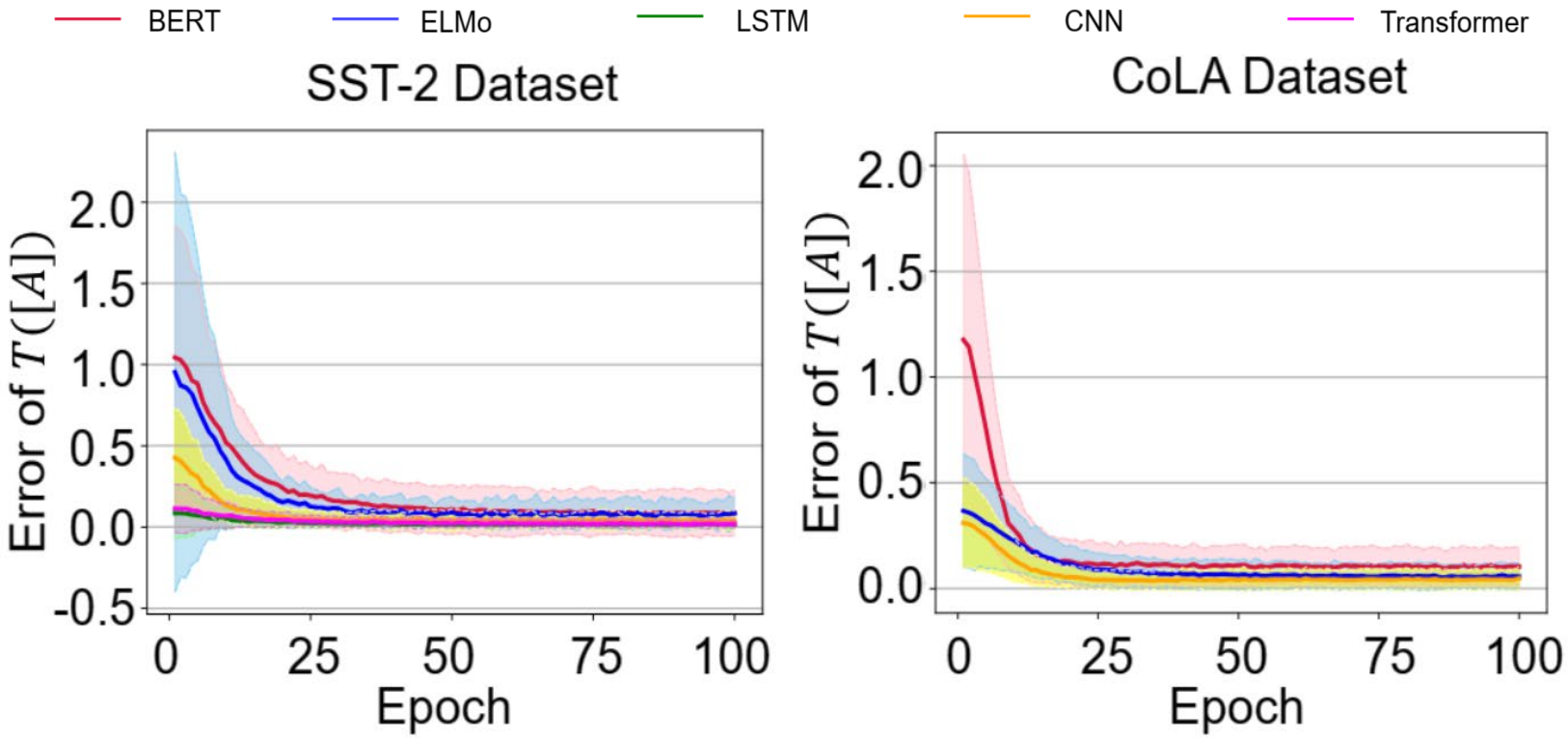
$$x_0|x_1|x_2|(x_3\&x_4)|(x_5\&x_6\&x_7)|x_8|(x_9\&x_{10}\&x_{11})$$

p_6 p_7 p_8

$$x_0|x_1|x_2|(x_3\&x_4)|(x_5\&x_6\&x_7)|x_8|(x_9\&x_{10}\&x_{11})$$

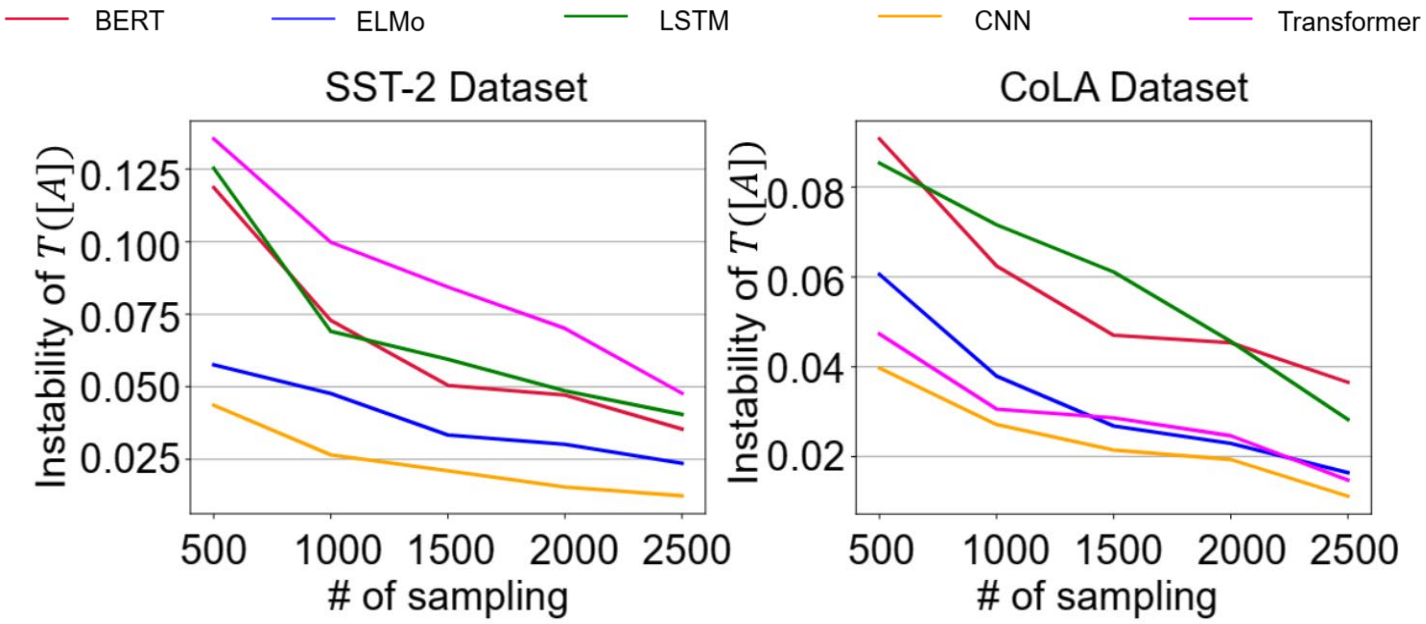
实验验证　近似的准确性

- 实验操作
  - 数据集：SST-2、CoLA； 模型：BERT、ELMo、LSTM、CNN、Transformer
  - 随机选择一组连续的单词A，衡量交互显著性$T([A])$的准确性：暴力枚举



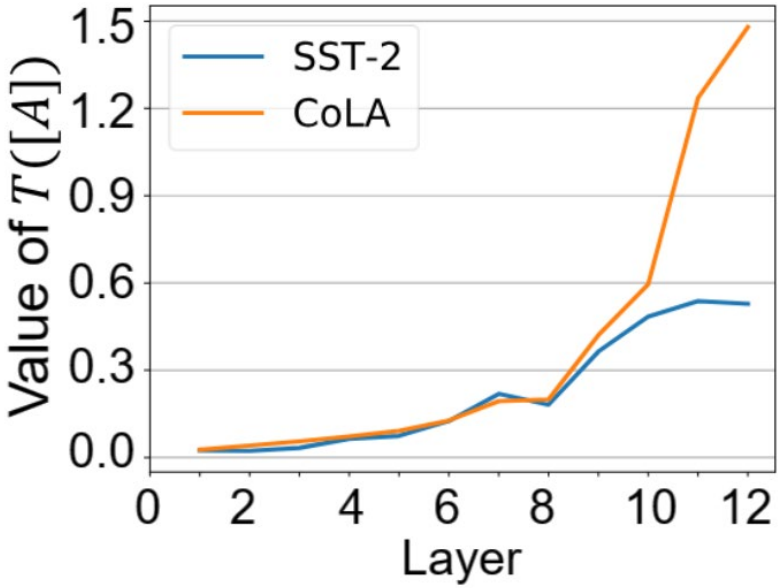
实验验证 $T([A])$的稳定性

- 实验操作
  - 使用g和S的不同采样集多次计算$T([A])$
  - 构造指标$\mathbb{E}_I[\frac{\mathbb{E}_{u,v:u\neq v}|T_{(u)}([A])-T_{(v)}([A])|}{\mathbb{E}_{\omega}|T_{(\omega)}([A])|}]$， $T_{(u)}([A])$表示第u个计算结果
- 结果
  - CNN的交互收敛的最快，不稳定性随采样次数增加而降低



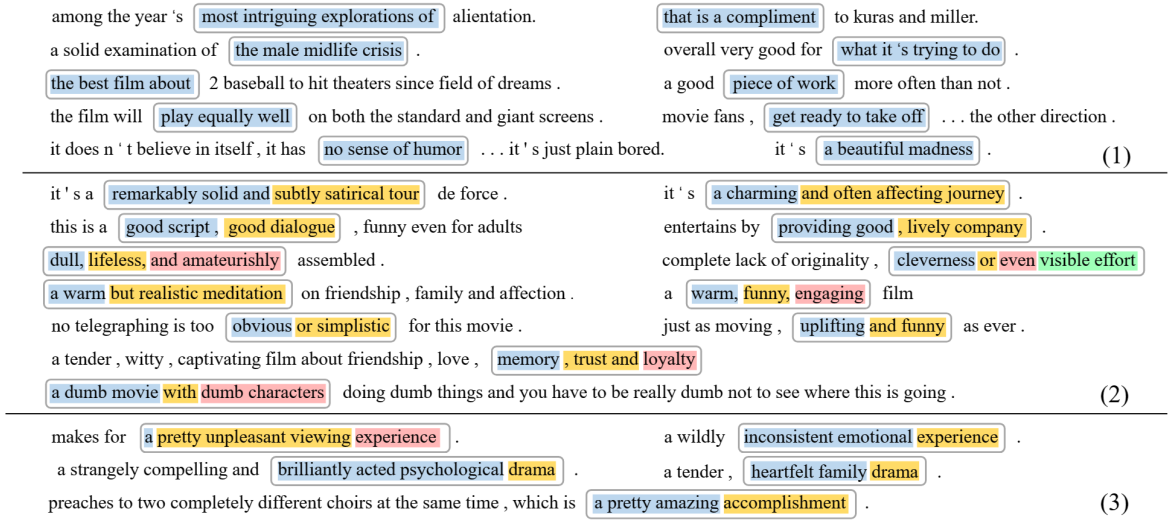
实验验证　中间层的交互

- 实验操作
  - 问题：中间层是高维向量，不是标量值
  - 构造指标$v_f(S)=\frac{\langle f_N, f_S\rangle}{\|f_N\|_2}$，　f_N表示由输入集合N获得的中间特征
  - 使用不同层的中间特征计算上述指标
- 结果
  - 输入变量之间的交互显著性是逐层增加的，原型特征在DNN深层构建



实验验证　DNN建模的原型特征

- 结果
  - 句子中的整个成分（短句或名词短语），BERT通常将整个成分视为单一的联盟，即原型特征
  - 包含连词或标点的单词集（and、or、”,”）等， BERT在连词或标点处划分不同的联盟（原型特征）
  - 由多个形容词和副词修饰的成分， BERT通常将所有形容词和状语合并为一个联盟，并将修饰后的成分作为另一个联盟



这些现象符合人类知识的句法结构解析

实验验证　洞察DNN的错误

- 实验操作
  - 目标：针对DNN的错误预测挖掘原型特征
  - 最大化和最小化每个句子的联盟的Shapley值
  - 例子：紧随其后的是一部同样忧郁、毫无幽默感的恐怖电影，尽管有缺陷，但它直白的恐怖手法值得称赞

maximum (prototypes towards incorrect predictions): if steven soderbergh ' s ' solaris ' is a failure it is a glorious failure . predict: negative

minimum (prototypes towards correct predictions): if steven soderbergh ' s ' solaris ' is a failure it is a glorious failure . label: positive

maximum (prototypes towards incorrect predictions): the longer the movie goes , the worse it gets , but it ' s actually pretty good in the first few minutes. predict: positive

minimum (prototypes towards correct predictions): the longer the movie goes , the worse it gets , but it ' s actually pretty good in the first few minutes. label: negative

maximum (prototypes towards incorrect predictions): on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . predict: negative

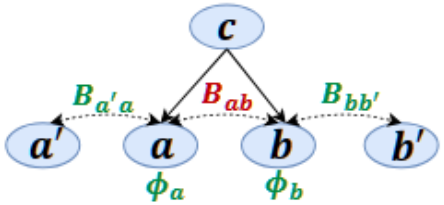
minimum (prototypes towards correct predictions): on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . label: positive

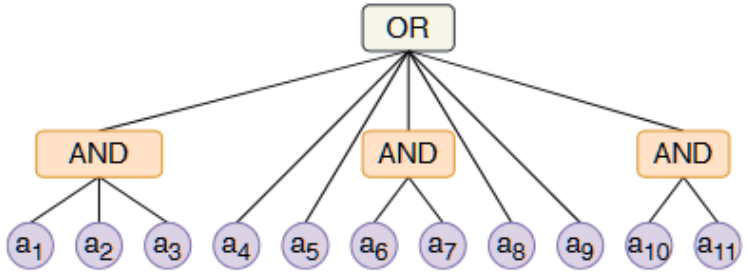
maximum (prototypes towards incorrect predictions): on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . predict: negative

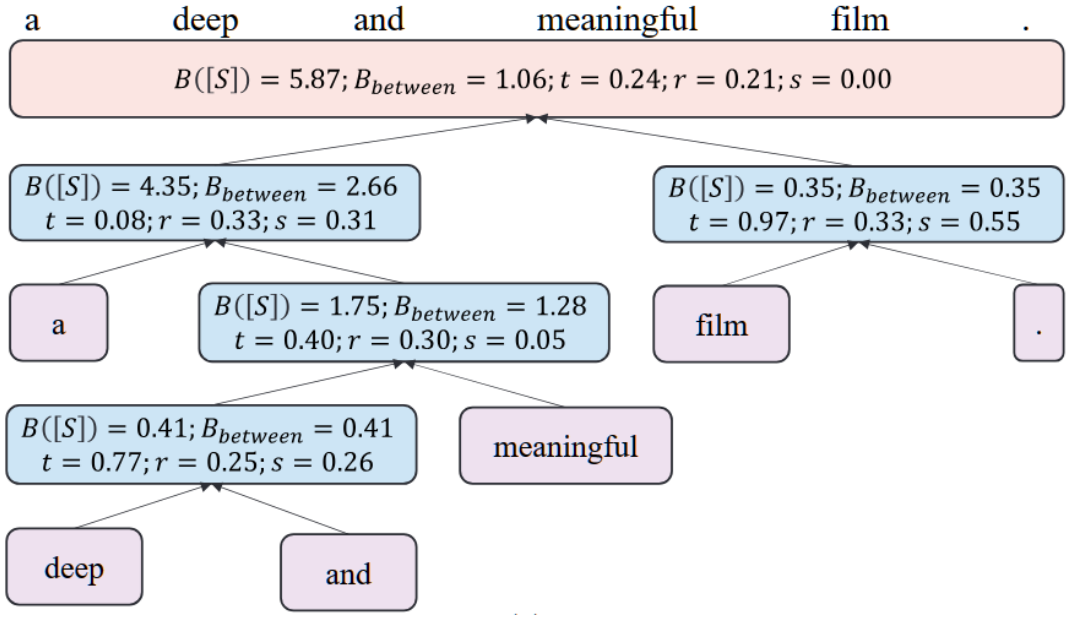
minimum (prototypes towards correct predictions): on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . label: positive

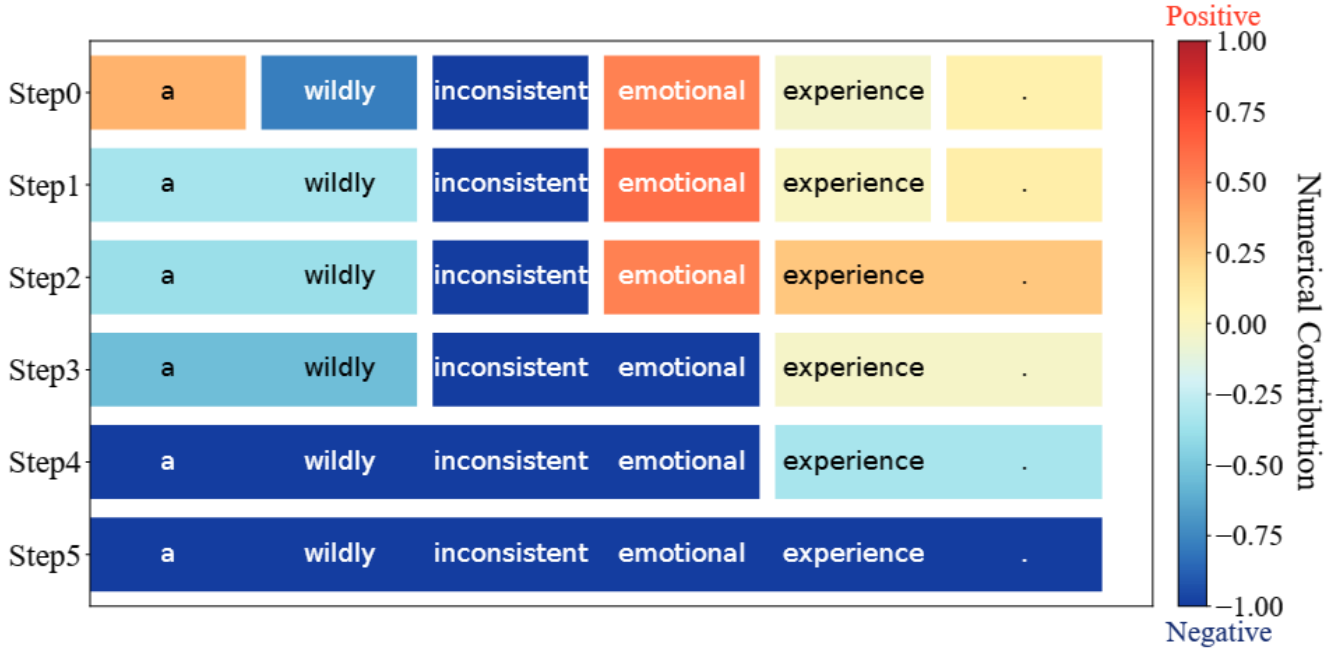
拓展实验 自然语言交互树

- 自然语言的潜在结构是树状的



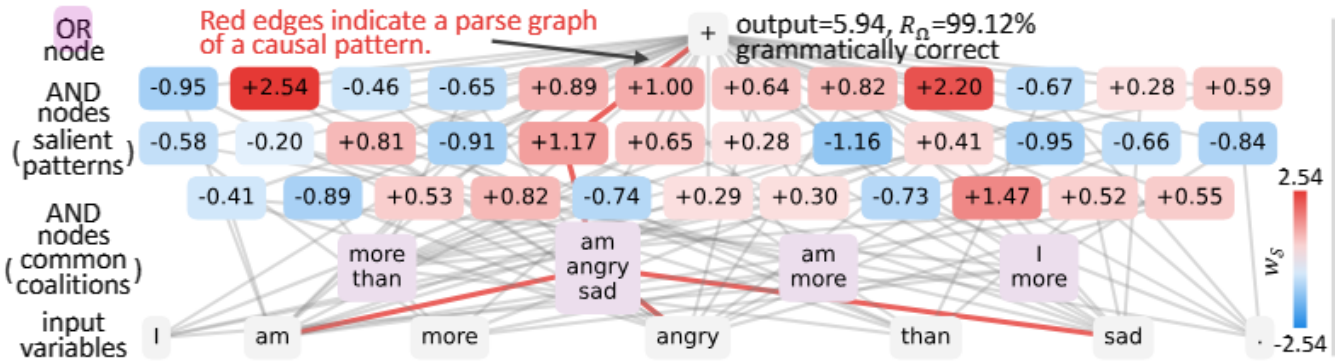


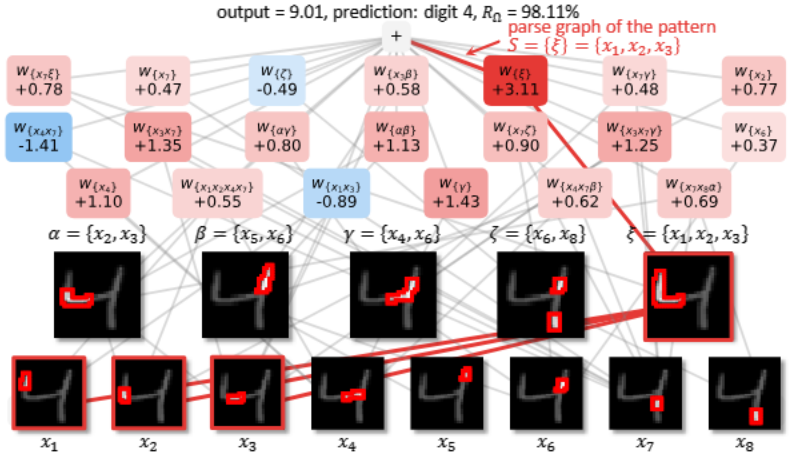




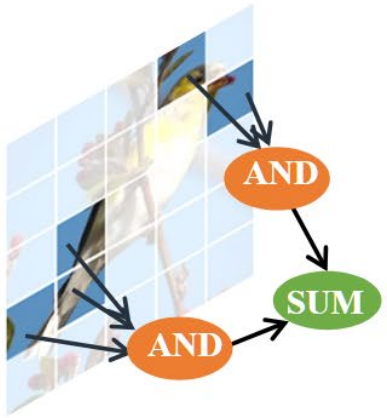
拓展工作

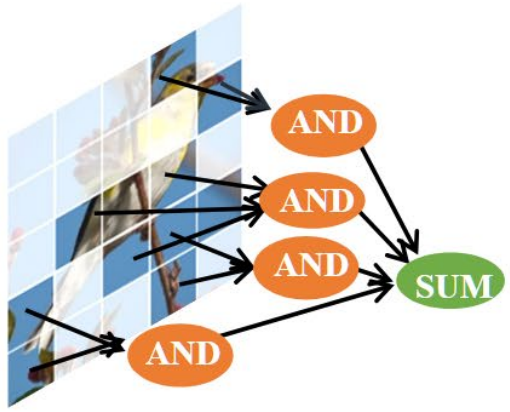
- 与或图

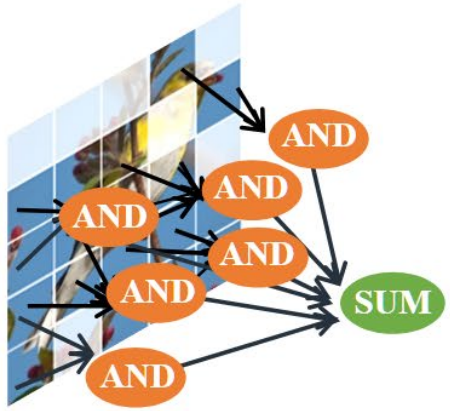




- 多阶交互（阶可理解为交互的复杂性）

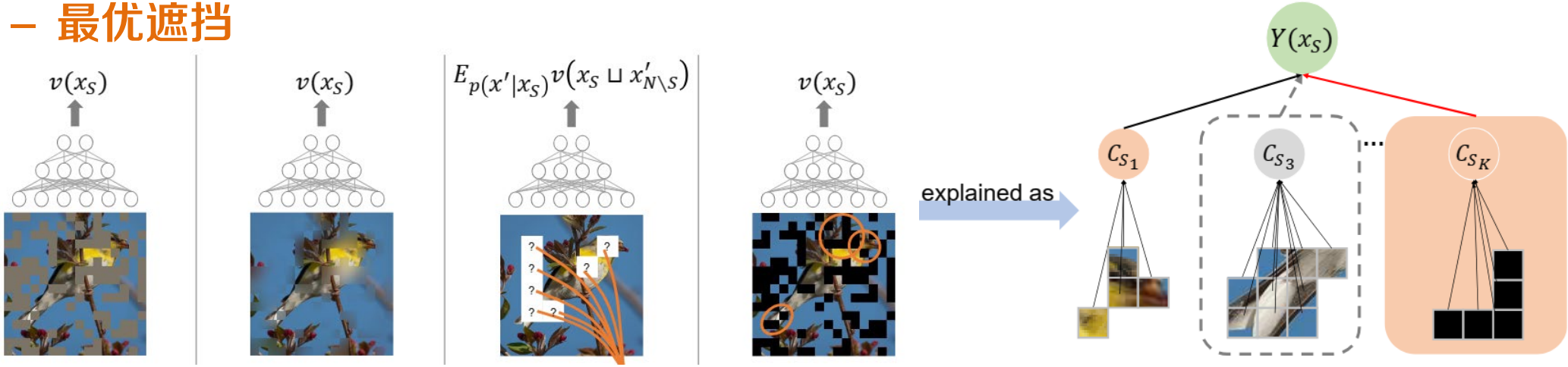






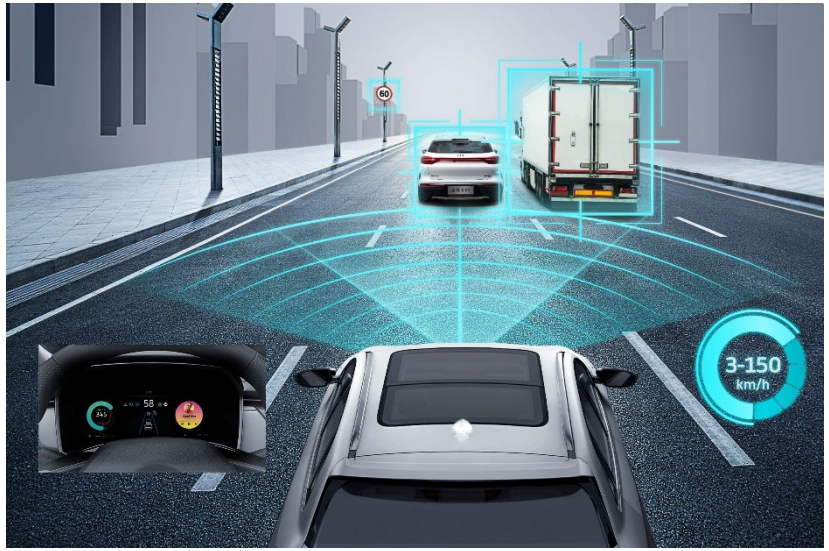
进阶思考

- 进阶思考
  - DNN究竟是否可以被解释（严格地解释为符号化概念）
    - 证明DNN是符号化的（无法严谨符号化注定语义层面的解释不是根本的）
  - 进一步解释
    - 证明交互概念的泛化性：同一交互概念在不同样本上的泛化性、同一交互概念在不同模型上的迁移性
    - 解释已有模型及前人可解释理论
  - 最优遮挡



特点总结与工作展望

- 特点
  - 基于博弈交互的统一体系，可解释性的理论基建
  - 经验性有偏方法转变为理论性无偏估计
  - 洞察模型建模知识与人类知识差异
  - 解释过往（共性特征）、优化未来（反馈指导）
- 工作展望
  - 解释错误
  - 提升性能（惩罚交互、不再端到端）
    - 大模型
    - 车载系统
  - 语义层面上与DNN的内在表达交流



参考文献

[1] Zhang H, et al. Interpreting Multivariate Shapley Interactions in DNNs[J]. AAAI Conference on Artificial Intelligence, 2021.

[2] Zhang H, et al. Game-Theoretic Interactions of Different Orders[J]. ArXiv:2010.14978, 2020.

[3] Zhang D, et al. Building Interpretable Interaction Trees for Deep NLP Models[J]. AAAI Conference on Artificial Intelligence, 2021.

[4] Ren J, et al. Defining and Quantifying the Emergence of Sparse Concepts in DNNs[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 20280-20289.

[5] Ren J, et al. Can We Faithfully Represent Masked States to Compute Shapley Values on a DNN?[J]. International Conference on Learning Representations (ICLR), 2023.

[6] Li M, Zhang Q. Does a neural network really encode symbolic concepts?[J]. International Conference on Machine Learning (ICML), 2023, 202:20452–20469.

[7] Deng H, et al. Discovering and Explaining the Representation Bottleneck of DNNs[J]. International Conference on Learning Representations (ICLR), 2022.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！