



对话系统开放意图检测与发现

分享者: 张瀚镭 (直博一年级在读)

导师: 徐华副教授

单位: 清华大学计算机系智能技术与系统国家重点实验室

2021年3月12日

个人主页: <https://hanleizhang.github.io>

目录



- ◎ 研究背景及现状
- ◎ 基于自适应决策边界的开放意图分类模型 (AAAI 2021录用)
- ◎ 基于深度对齐聚类的新意图发现模型 (AAAI 2021录用)
- ◎ 总结和展望
- ◎ 附录



目录



- ◎ 研究背景及现状
- ◎ 基于自适应决策边界的开放意图分类模型 (AAAI 2021录用)
- ◎ 基于深度对齐聚类的新意图发现模型 (AAAI 2021录用)
- ◎ 总结和展望
- ◎ 附录





研究背景

- ◎ 人机交互技术迅猛发展，智能聊天机器人需求广泛
 - ◆ 智能客服：食品饮料、医疗保健、电子商务、电信运营
 - ◆ 休闲娱乐：智能音箱、语音助手（微软小冰、苹果Siri、百度小度、小爱同学等等）
- ◎ 用户需求复杂多样，对话系统难以识别全部意图
 - ◆ 检测并发现用户开放意图并加以有效利用具有很大的商业前景
- ◎ 发现用户对对话开放意图面临的困难
 - ◆ 如何将开放意图与已知意图分离？
 - ◆ 如何发现开放意图的细粒度类别？



对话系统无法处理的开放意图



语音助手





研究背景

◎ 开放意图检测

- ◆ 开放域分类问题
- ◆ N类已知意图，开放意图全部归为一类
- ◆ 训练集只包含已知意图

用户说话内容	意图标签
我想订明天去北京的机票。	订票
附近有没有卖披萨的地方？	点餐
今天晚上有没有去美国的航班？	订票
我想去附近一家中餐馆吃饭。	点餐
...	...
今天天气怎么样？	开放意图
你能帮我看看这个是什么吗？	开放意图

◎ 开放意图发现

- ◆ 半监督聚类问题
- ◆ N类已知意图，M类细粒度开放意图
- ◆ 训练集包含少量有标注已知意图+大量未标注意图

训练集		测试集
已知意图 1	有标注	意图 1
已知意图 2	有标注	意图 2
...	有标注	...
已知意图 N	有标注	意图 N
开放意图 1		意图 N+1
开放意图 2		意图 N+2
...		...
开放意图 M		意图 N+M



研究现状

◎ 开放意图检测

◆ 开放域分类

- 基于SVM方法: [Scheirer et.al, 2013; Liu and Fei, 2016]
- 基于深度神经网络方法: [Bendale and Boulton, 2016; Shu et.al, 2017]

◆ 未知意图检测

- 基于意图特征密度: [Lin and Xu, 2019; Yan, Fan and Li 2020]

◎ 开放意图发现

◆ 基于无监督聚类的方法

- 特征工程: [Padmasundari and Srinivas, 2018; Shi et.al, 2018] 语义解析器: [Hakkani-Tür et.al, 2018]

◆ 基于半监督聚类的方法

- 数据对相似性: [Hsu et al. 2018, 2019; Lin et al. 2020] 迁移学习: [Han et.al, 2019]





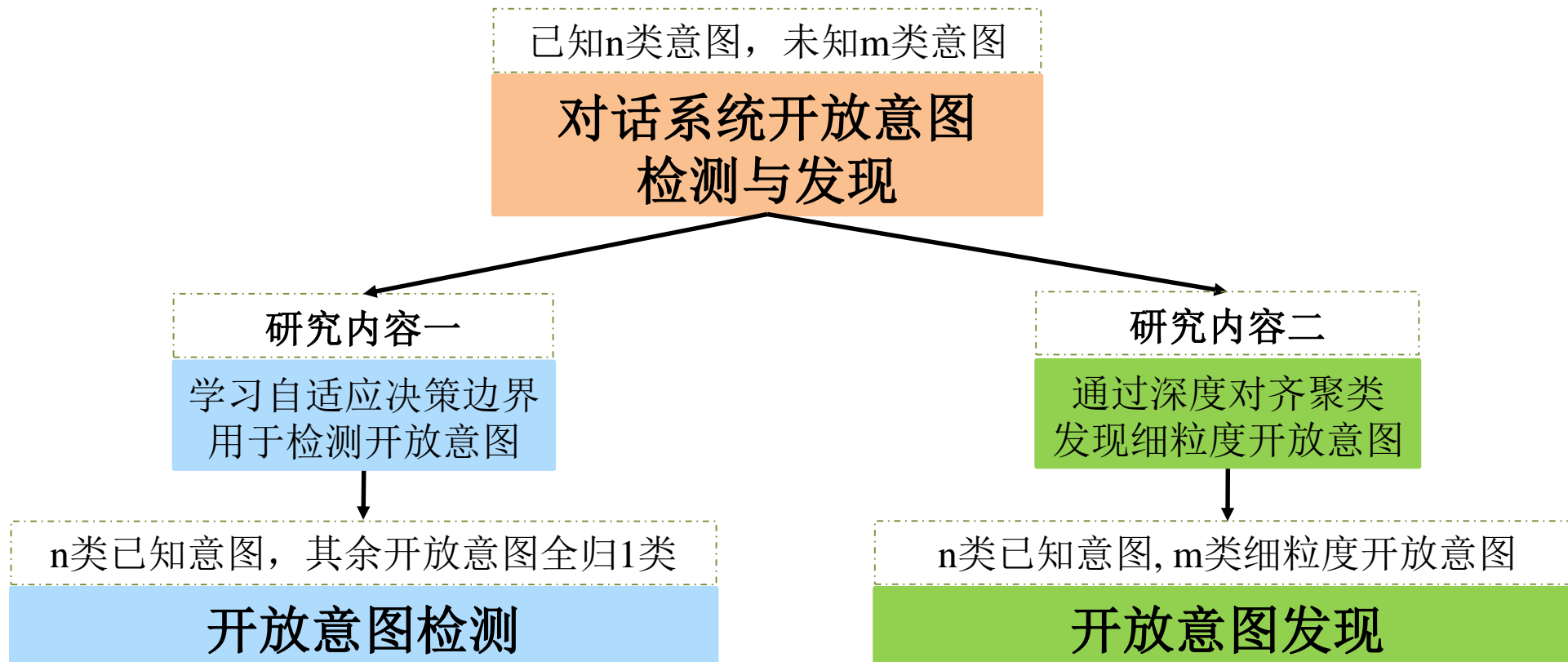
研究现状

现有研究方法	存在的不足	本文研究思路	核心思想
开放域分类	需要开放域样本用于训练 决策阈值需要人工筛选	基于自适应决策边界的 开放意图分类模型	基于预训练特征表示 自适应学习决策边界 的后处理方法
未知意图检测	需要设计专门分类器 模型性能受超参影响大		
基于无监督聚类	在缺乏先验知识引导下， 难以获得满意聚类结果。	基于深度对齐聚类的新 意图发现模型	利用少量有标注意力图 作为先验知识 通过簇中心对齐构建 一致性自监督信号
基于半监督聚类	先验知识泛化性差，导致 模型过拟合。无监督数据 缺乏高质量自监督信号。		





研究框架



目录



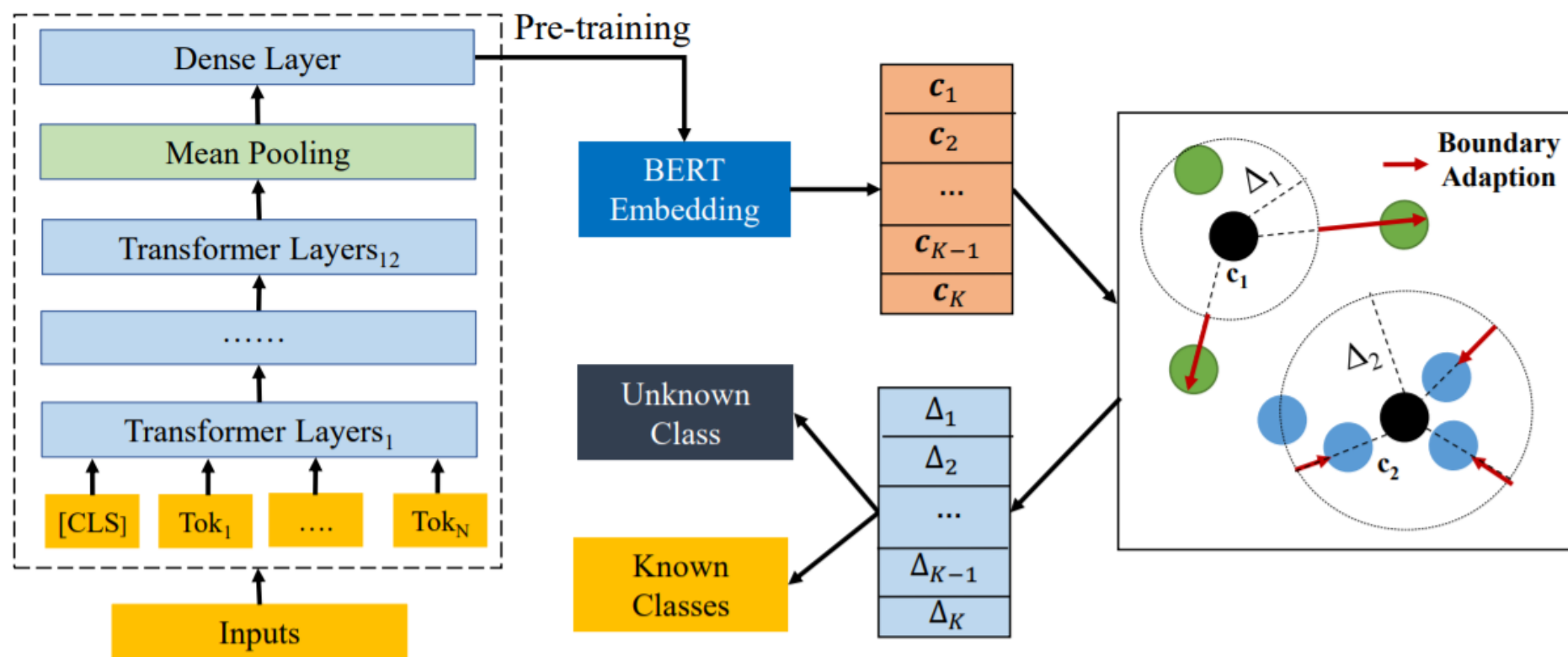
- 研究背景及现状
- 基于自适应决策边界的开放意图分类模型 (AAAI 2021录用)**
- 基于深度对齐聚类的新意图发现模型 (AAAI 2021录用)
- 总结和展望
- 附录





基于自适应决策边界的开放意图分类方法 (ADB)

- 基于预训练意图特征学习自适应决策边界检测开放意图



意图特征表示

簇中心和决策边界定义

决策边界自适应学习





意图特征表示

利用BERT抽取深度意图表示 [Devlin et.al, 2019]

- 输入BERT第 i 句话 \mathbf{s}_i ，经过12层Transformer编码层

- 获得词向量表示 $[C, T_1, \dots, T_N] \in \mathbb{R}^{(N+1) \times H}$

- 平均池化操作得到句向量表示 $\mathbf{x}_i \in \mathbb{R}^H$

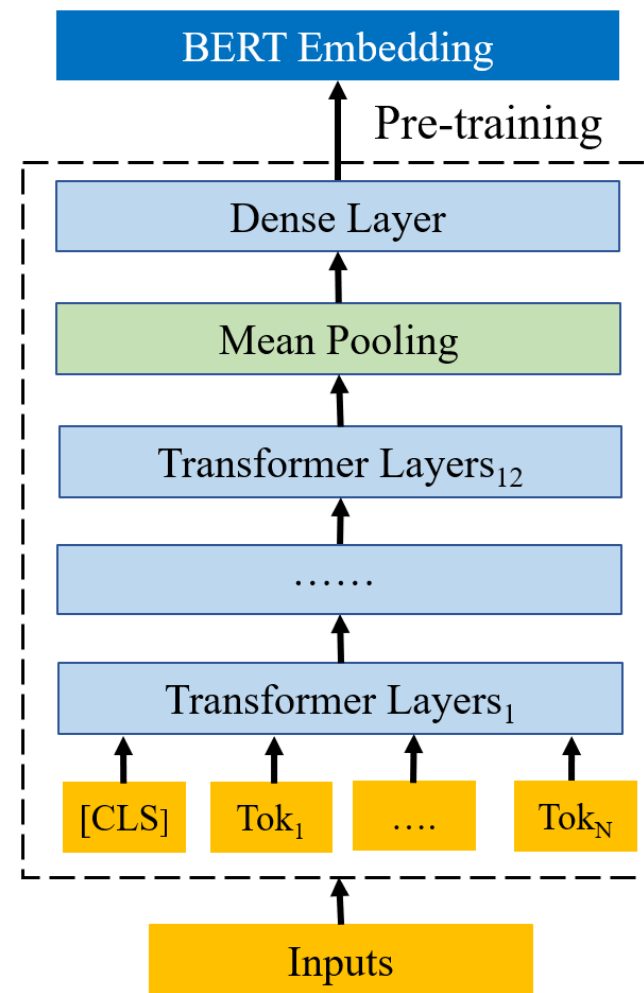
$$\mathbf{x}_i = \text{mean-pooling}([C, T_1, \dots, T_N])$$

- 经过致密层 h 得到意图特征表示 $\mathbf{z}_i \in \mathbb{R}^D$

$$\mathbf{z}_i = h(\mathbf{x}_i) = \sigma(W_h \mathbf{x}_i + b_h)$$

- 利用已知意图softmax分类进行预训练

$$\mathcal{L}_s = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\phi(\mathbf{z}_i)^{y_i})}{\sum_{j=1}^K \exp(\phi(\mathbf{z}_i)^j)}$$





簇中心和决策边界

◎ 定义簇中心和决策边界

- ◆ 对于第 k 类意图表示集合 S_k , 计算簇中心 $\mathbf{c}_k \in \mathbb{R}^D$

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(z_i, y_i) \in S_k} z_i$$

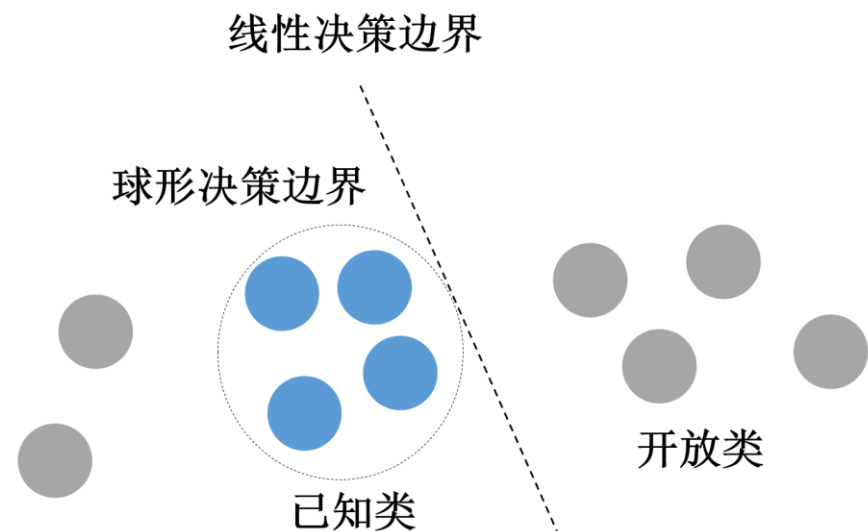
- ◆ 对于每类已知意图定义球形决策边界 Δ_k

- 希望满足目标:

$$\forall z_i \in S_k, \|z_i - \mathbf{c}_k\|_2 \leq \Delta_k$$

- 利用SoftPlus激活函数学习决策边界 Δ_k

$$\Delta_k = \log \left(1 + e^{\widehat{\Delta}_k} \right)$$





自适应决策边界

自适应决策边界学习

定义边界损失学习紧凑的决策边界

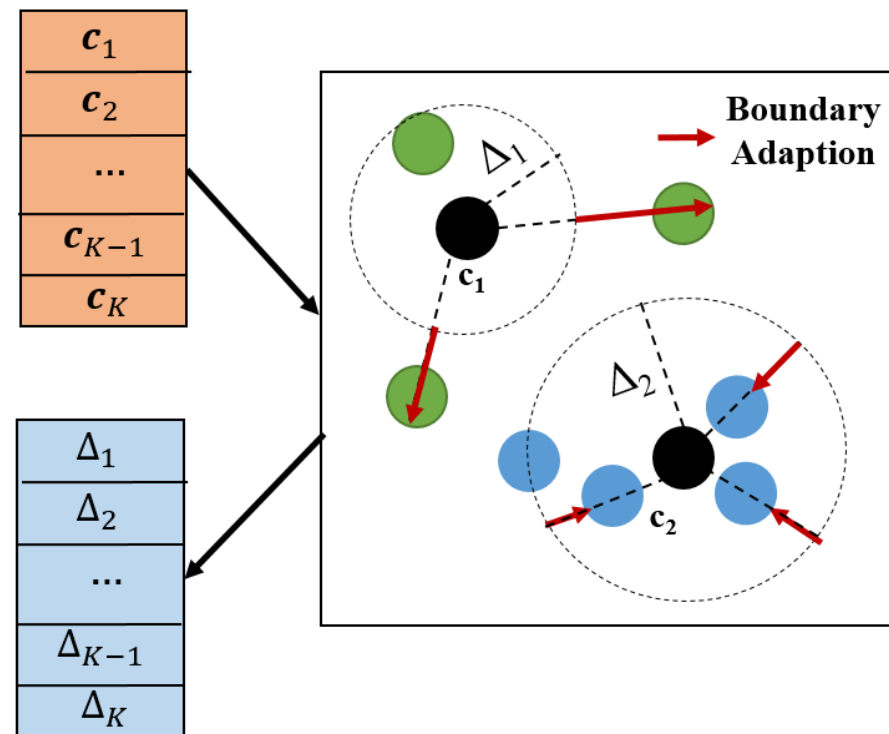
- 决策边界需要包含多数已知意图 (经验风险)
- 决策边界距离其对应的簇中心不能太远 (开放空间风险)

$$\mathcal{L}_b = \frac{1}{N} \sum_{i=1}^N [\delta_i (\|z_i - c_{y_i}\|_2 - \Delta_{y_i}) + (1 - \delta_i) (\Delta_{y_i} - \|z_i - c_{y_i}\|_2)]$$

$$\delta_i := \begin{cases} 1, & \text{if } \|z_i - c_{y_i}\|_2 > \Delta_{y_i}, \\ 0, & \text{if } \|z_i - c_{y_i}\|_2 \leq \Delta_{y_i}. \end{cases}$$

优化边界参数

$$\widehat{\Delta}_k := \widehat{\Delta}_k - \eta \frac{\partial \mathcal{L}_b}{\partial \widehat{\Delta}_k} \quad \frac{\partial \mathcal{L}_b}{\partial \widehat{\Delta}_k} = \frac{\sum_{i=1}^N \delta' (y_i = k) \cdot (-1)^{\delta_i}}{\sum_{i=1}^N \delta' (y_i = k)} \cdot \frac{1}{1 + e^{-\widehat{\Delta}_k}}$$





实验设置

实验数据集

- ◆ BANKING银行交易数据集 [Casanueva et al. 2020]
- ◆ OOS开放域对话数据集 [Larson et.al, 2019]
- ◆ StackOverflow技术问答标题数据集 [Xu et.al, 2015]

数据集	类别数	词表大小	训练集	验证集	测试集
BANKING	77	5,028	9,003	1,000	3,080
OOS	150	8,376	15,000	3,000	5,700
StackOverflow	20	17,182	12,000	2,000	6,000

训练和测试

- ◆ 训练集只包含已知意图，已知意图比例设置为25%，50%和75%
- ◆ 测试集包含全部意图，将除已知意图外其他意图看作一类开放意图





实验设置

◎ 评价指标

- ◆ 总体性能: Macro F1-score, Accuracy (全部类别)
- ◆ 细粒度性能: Macro F1-score (在已知类和开放类别)

	参数设置
句子最大长度	BANKING: 55, OOS: 30, StackOverflow: 45
意图特征维度	768
边界学习率	0.05
学习率	0.00002
最大训练迭代次数	100
批处理大小	128
意图特征维度	768
边界学习率	0.05





实验结果

- 在三个数据集上，将已知意图比例设置为25%，50%和75%
 - 我们提出的方法ADB在9个设置中，全部取得当前最佳结果 (2%~40%的性能提升)

	Methods	BANKING		OOS		StackOverflow	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
25%	MSP	43.67	50.09	47.02	47.62	28.67	37.85
	DOC	56.99	58.03	74.97	66.37	42.74	47.73
	OpenMax	49.94	54.14	68.50	61.99	40.28	45.98
	DeepUnk	64.21	61.36	81.43	71.16	47.84	52.05
	ADB	78.85	71.62	87.59	77.19	86.72	80.83
50%	MSP	59.73	71.18	62.96	70.41	52.42	63.01
	DOC	64.81	73.12	77.16	78.26	52.53	62.84
	OpenMax	65.31	74.24	80.11	80.56	60.35	68.18
	DeepUnk	72.73	77.53	83.35	82.16	58.98	68.01
	ADB	78.86	80.90	86.54	85.05	86.40	85.83
75%	MSP	75.89	83.60	74.07	82.38	72.17	77.95
	DOC	76.77	83.34	78.73	83.59	68.91	75.06
	OpenMax	77.45	84.07	76.80	73.16	74.42	79.78
	DeepUnk	78.52	84.31	83.71	86.23	72.33	78.28
	ADB	81.08	85.96	86.32	88.53	82.78	85.99





实验结果-辅助实验 1

在已知类和开放类分别进行评估

我们提出的方法ADB在9个设置中，同样全部取得当前最佳结果

	Methods	BANKING		OOS		StackOverflow	
		Unknown	Known	Unknown	Known	Unknown	Known
25%	MSP	41.43	50.55	50.88	47.53	13.03	42.82
	DOC	61.42	57.85	81.98	65.96	41.25	49.02
	OpenMax	51.32	54.28	75.76	61.62	36.41	47.89
	DeepUnk	70.44	60.88	87.33	70.73	49.29	52.60
	ADB	84.56	70.94	91.84	76.80	90.88	78.82
50%	MSP	41.19	71.97	57.62	70.58	23.99	66.91
	DOC	55.14	73.59	79.00	78.25	25.44	66.58
	OpenMax	54.33	74.76	81.89	80.54	45.00	70.49
	DeepUnk	69.53	77.74	85.85	82.11	43.01	70.51
	ADB	78.44	80.96	88.65	85.00	87.34	85.68
75%	MSP	39.23	84.36	59.08	82.59	33.96	80.88
	DOC	50.60	83.91	72.87	83.69	16.76	78.95
	OpenMax	50.85	84.64	76.35	73.13	44.87	82.11
	DeepUnk	58.54	84.75	81.15	86.27	37.59	81.00
	ADB	66.47	86.29	83.92	88.58	73.86	86.80

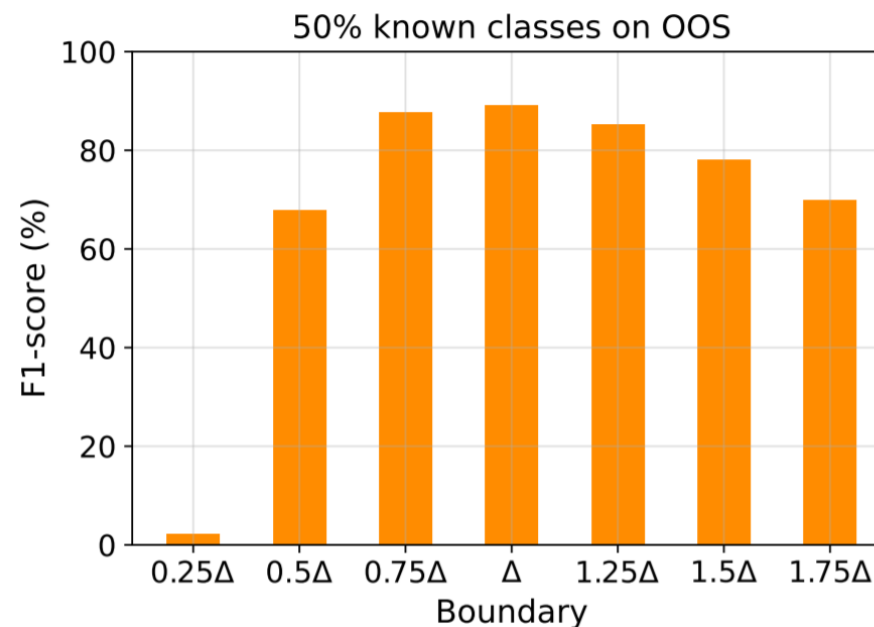
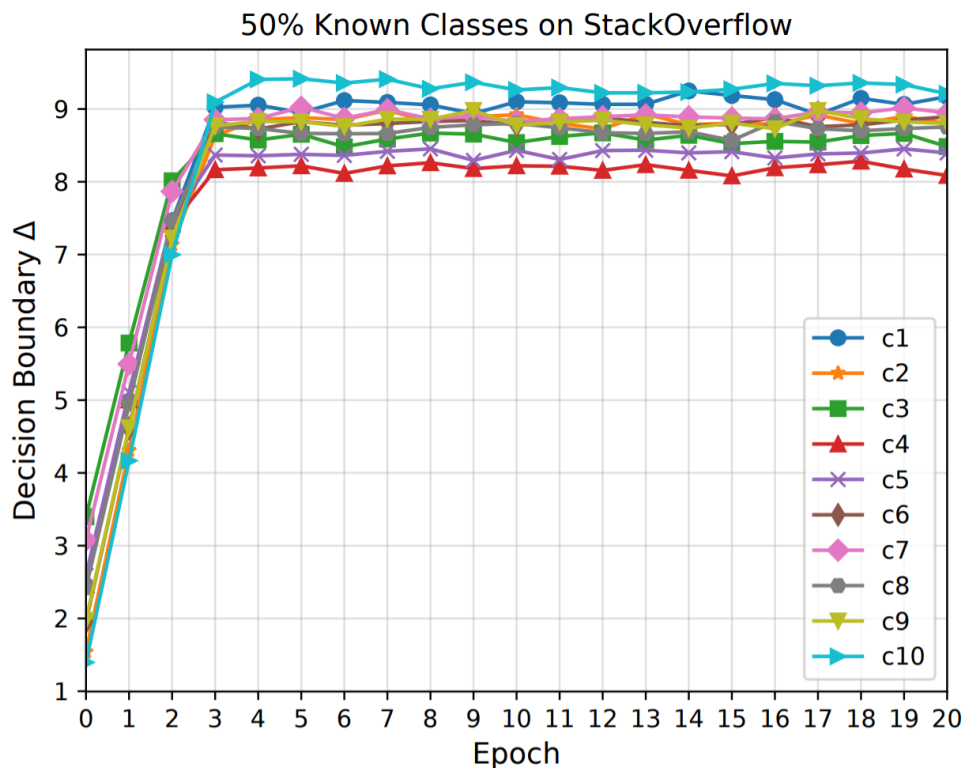




实验结果-辅助实验 2

决策边界学习过程及影响

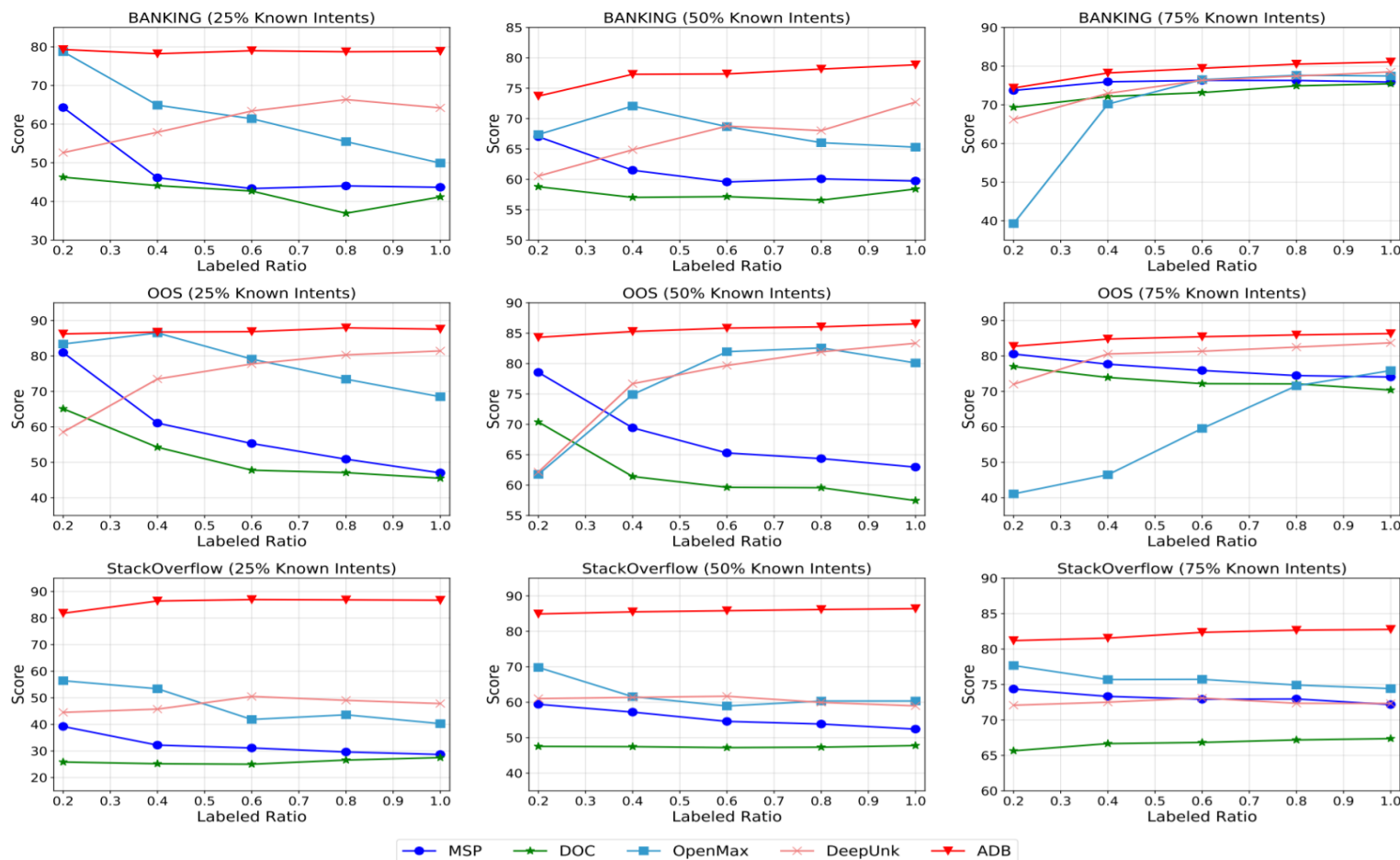
决策边界学习过程 (左图) 以及 决策边界紧凑程度影响 (右图)





实验结果-辅助实验 3

有标注数据比例对实验结果的影响 (已知意图比例25%, 50%和75%)





目录



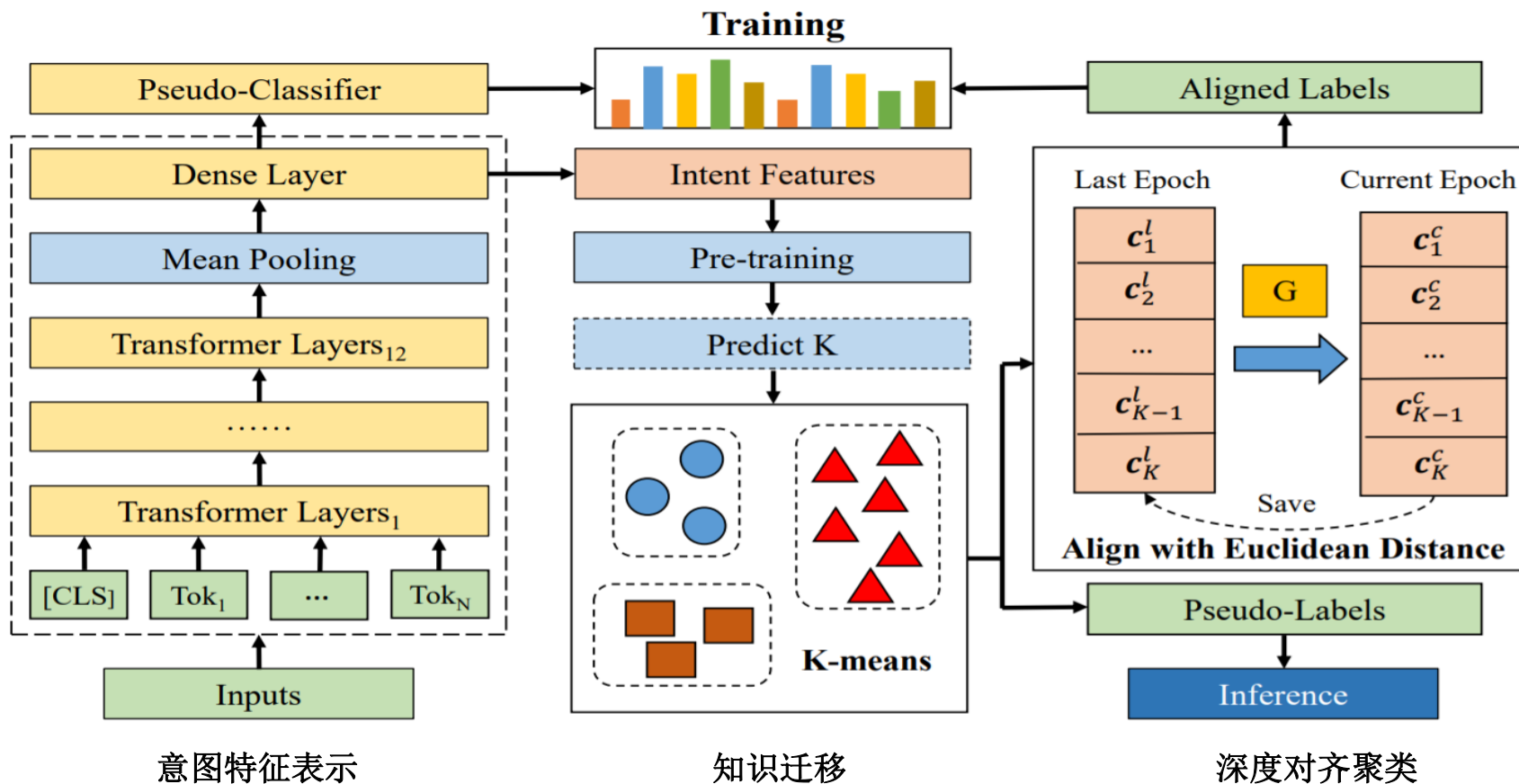
- 研究背景及现状
- 基于自适应决策边界的开放意图分类模型 (AAAI 2021 录用)
- 基于深度对齐聚类的新意图发现模型 (AAAI 2021 录用)**
- 总结和展望
- 附录





基于深度对齐聚类的开放意图发现 (DeepAligned)

- 利用先验知识和自监督信号指导聚类发现开放意图



意图特征表示

知识迁移

深度对齐聚类



基于深度对齐聚类的开放意图发现 (DeepAligned)



◎ 知识迁移

- ◆ 利用BERT特征提取器获得意图表示 (方法同开放意图检测)
- ◆ 利用少量有标注数据进行预训练
 - Softmax监督分类
- ◆ 将预训练分类层去掉, 保留剩余部分参数作为先验知识
- ◆ 利用经过初始化的意图特征预测簇个数 K
 - 设定较大聚类簇个数 K' 进行K-Means聚类
 - 通过保留高置信度簇 (簇个数大于阈值 t) 估算真实簇个数

$$K = \sum_{i=1}^{K'} \delta(|S_i| \geq t)$$



基于深度对齐聚类的开放意图发现 (DeepAligned)



无监督聚类和表示学习 DeepCluster [Caron et.al, 2018]

◆ 无监督聚类

- 利用K-Means聚类产生簇分配 $\{y_i\}_{i=1}^N$ 和簇中心矩阵 C

$$\min_{C \in \mathbb{R}^{K \times D}} \frac{1}{N} \sum_{i=1}^N \min_{y_i \in \{1, \dots, K\}} \|I_i - C_{y_i}\|_2^2$$

◆ 表示学习

- 利用聚类产生的簇分配作为伪标签 \rightarrow 监督信号

◆ 存在的问题

- 每轮迭代簇分配序号随机 \rightarrow 伪标签不一致
- 需要重新初始化分类层参数，无法保留历史训练信息





基于深度对齐聚类的开放意图发现 (DeepAligned)

深度对齐聚类 DeepAligned

伪标签不一致性问题

- 发现: DeepCluster未利用聚类产生的簇中心矩阵 (产生簇分配的目标)
- 解决方案: 通过相邻迭代簇中心对齐获得一致性伪标签

对齐策略

- 匈牙利算法获得欧式空间簇中心对齐映射 $G: C^c = G(C^l)$
- 利用 G 的逆映射 G^{-1} 作用在当前伪标签 y^c 产生对齐伪标签 $y^{align}: y^{align} = G^{-1}(y^c)$

- 利用伪标签进行自监督学习:
$$\mathcal{L}_s = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\phi(\mathbf{I}_i)^{y_i^{align}})}{\sum_{j=1}^K \exp(\phi(\mathbf{I}_i)^j)}$$

- 轮廓系数评价无监督聚类簇质量:
$$SC = \frac{1}{N} \sum_{i=1}^N \frac{b(\mathbf{I}_i) - a(\mathbf{I}_i)}{\max\{a(\mathbf{I}_i), b(\mathbf{I}_i)\}}$$





实验设置

实验数据集

- ◆ BANKING银行交易数据集 [Casanueva et al. 2020]
- ◆ CLINC意图分类数据集 [Larson et.al, 2019]

数据集	类别数	词表大小	训练集	验证集	测试集
BANKING	77	5,028	9,003	1,000	3,080
OOS	150	8,376	15,000	3,000	5,700

训练和测试 [Lin et.al, 2020]

- ◆ 训练集包含10%有标注数据，其余均为无标注数据
- ◆ 有标注数据只包含已知意图 (占比75%)，无标注数据包含已知和未知意图
- ◆ 测试集包含全部意图





实验设置

◎ 评价指标

- ◆ 聚类指标: 标准化互信息 NMI, 调整兰德指数 ARI, 准确率 ACC
- ◆ 利用匈牙利算法将预测簇序号与真实标签通过匈牙利算法对齐, 计算ACC

	参数设置
句子最大长度	BANKING: 55, CLINC: 30
意图特征维度	768
边界学习率	0.05
学习率	0.00005
最大训练迭代次数	100
批处理大小	128
意图特征维度	768





实验结果

- 7个无监督对比方法，5个半监督对比方法
 - 我们提出的方法DeepAligned，全部取得当前最佳结果 (3%~14%的性能提升)

	Method	CLINC			BANKING		
		NMI	ARI	ACC	NMI	ARI	ACC
Unsupervised.	KM	70.89	26.86	45.06	54.57	12.18	29.55
	AG	73.07	27.70	44.03	57.07	13.31	31.58
	SAE-KM	73.13	29.95	46.75	63.79	22.85	38.92
	DEC	74.83	27.46	46.89	67.78	27.21	41.29
	DCN	75.66	31.15	49.29	67.54	26.81	41.99
	DAC	78.40	40.49	55.94	47.35	14.24	27.41
	DeepCluster	65.58	19.11	35.70	41.77	8.95	20.69
Semi-supervised.	PCK-means	68.70	35.40	54.61	48.22	16.24	32.66
	BERT-KCL	86.82	58.79	68.86	75.21	46.72	60.15
	BERT-MCL	87.72	59.92	69.66	75.68	47.43	61.14
	CDAC+	86.65	54.33	69.89	72.25	40.97	53.83
	BERT-DTC	90.54	65.02	74.15	76.55	44.70	56.51
	DeepAligned	93.89	79.75	86.49	79.56	53.64	64.90





实验结果-辅助实验 1

◎ 消融实验和预测 K

◆ 是否利用有标注是否利用有标注数据预训练和对齐策略对实验结果的影响

	Method	CLINC			BANKING		
		NMI	ARI	ACC	NMI	ARI	ACC
Without Pre-training	Reinitialization	57.80	9.63	23.02	34.34	4.49	13.67
	Alignment	62.53	14.10	28.63	36.91	5.23	15.42
With Pre-training	Reinitialization	82.90	45.67	55.80	68.12	31.56	41.32
	Alignment	93.89	79.75	86.49	79.56	53.64	64.90

◆ 未知簇个数时 (指定为实际值两倍) 预测 K 的结果

	Method	CLINC (K'=300)		BANKING (K'=154)	
		K (Pred)	Error	K (Pred)	Error
25%	BERT-MCL	38	75.00	19	75.32
	BERT-DTC	94	37.33	37	51.95
	DeepAligned	122	18.67	66	14.29
50%	BERT-MCL	75	50.00	38	50.65
	BERT-DTC	131	12.67	71	7.79
	DeepAligned	130	13.33	64	16.88

	Methods	CLINC (K'=300)		BANKING (K'=154)	
		K (Pred)	Error	K (Pred)	Error
75%	BERT-MCL	112	25.33	58	24.68
	BERT-DTC	195	30.00	110	42.86
	DeepAligned	129	14.00	67	12.99

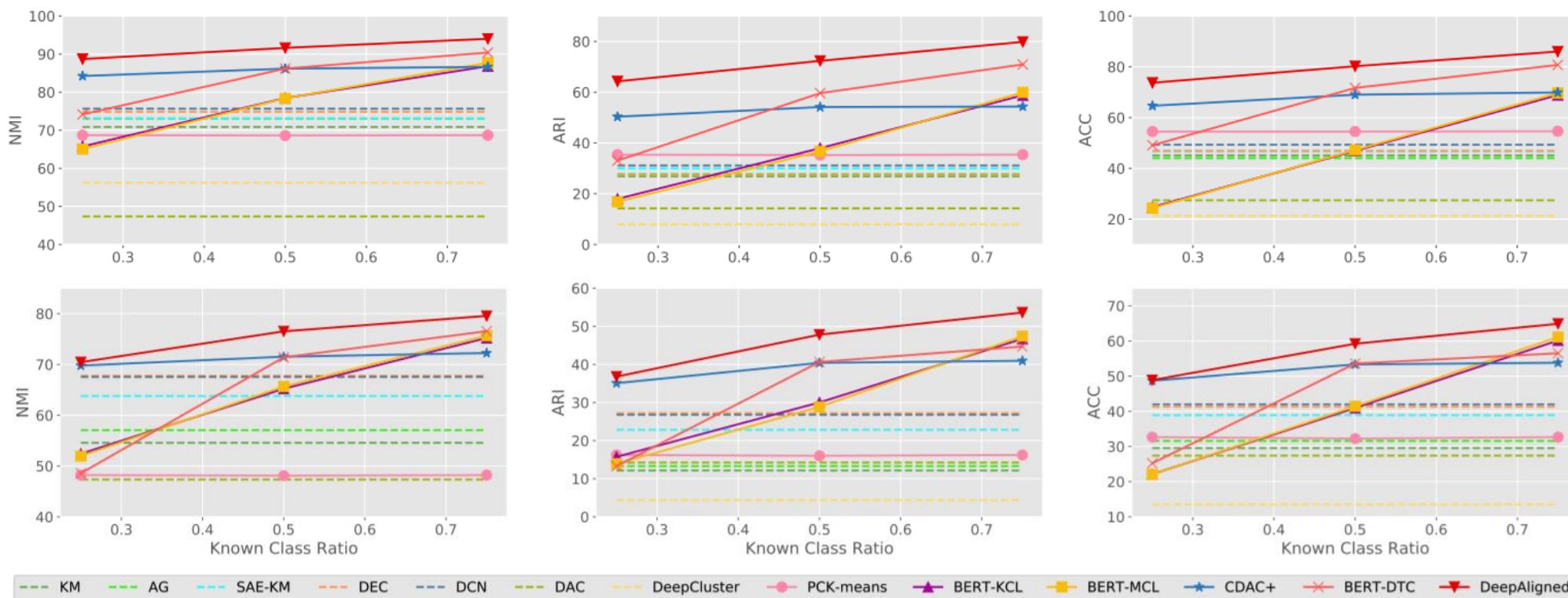




实验结果-辅助实验 2

已知意图比例对实验结果的影响

◆ CLINC数据集 (上图) 以及 BANKING数据集(下图)

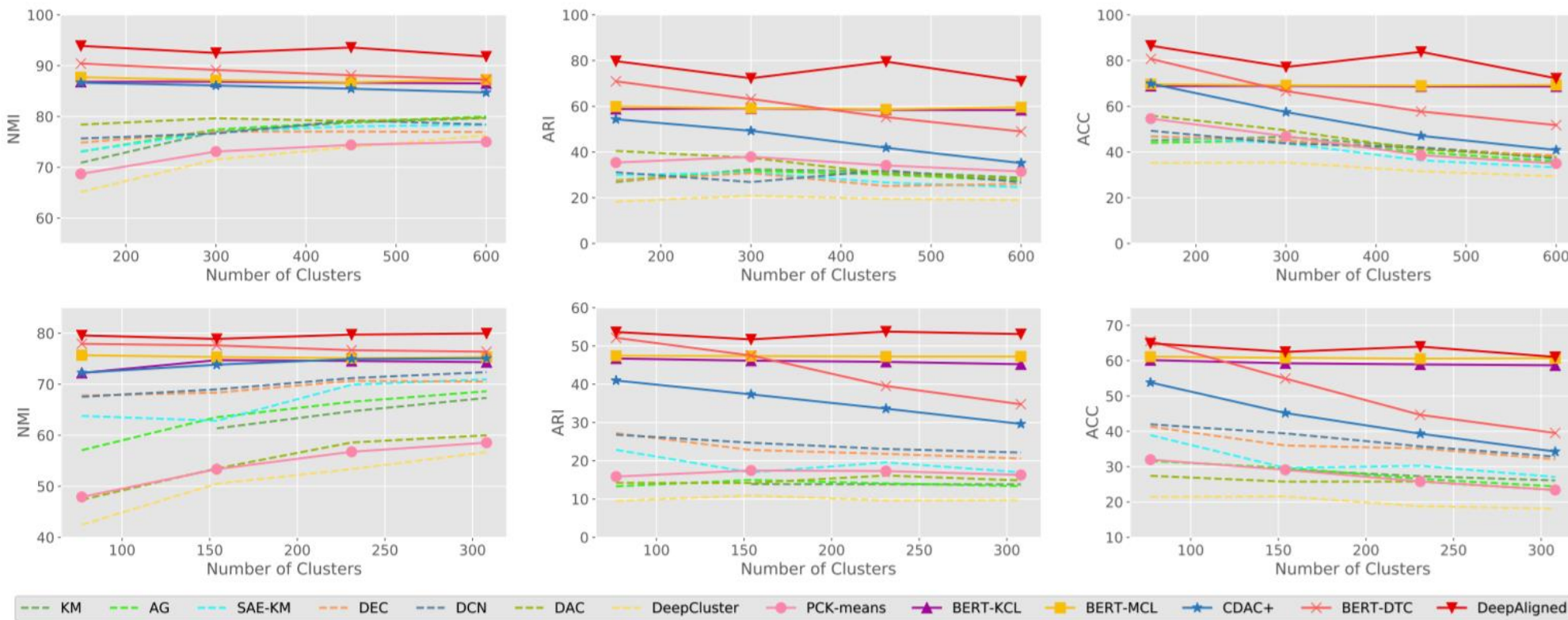




实验结果-辅助实验 3

有标注数据比例对实验结果的影响

◆ CLINC数据集 (上图)和BANKING数据集 (下图)





目录

- ◎ 研究背景及现状
- ◎ 基于自适应决策边界的开放意图分类模型 (AAAI 2021录用)
- ◎ 基于深度对齐聚类的新意图发现模型 (AAAI 2021录用)
- ◎ **总结和展望**
- ◎ 附录





总结和展望

◎ 基于自适应决策边界的开放意图分类模型

- ◆ 定义球形决策边界，通过平衡经验风险和开放空间风险学习**紧凑的决策边界**
- ◆ 无需开放域样本参与，自动学习决策边界**避免复杂调参**
- ◆ 允许任意神经网络分类器，在**不修改模型架构**的情况下检测开放意图

◎ 基于深度对齐聚类的开放意图发现模型

- ◆ 利用少量有标注数据进行**知识迁移**，指导聚类意图发现过程
- ◆ 通过对齐策略获得**高质量自监督信号**用于意图特征表示学习

◎ 未来工作

- ◆ 拓展到包含上下文的**多轮对话领域**
- ◆ 拓展到包含不同模态的**多模态领域**





目录

- ◎ 研究背景及现状
- ◎ 基于自适应决策边界的开放意图分类模型 (AAAI 2021录用)
- ◎ 基于深度对齐聚类的新意图发现模型 (AAAI 2021录用)
- ◎ 总结和展望
- ◎ 附录





亮点

- ◎ 我们在本次AAAI2021大会录用的两篇论文，欢迎大家关注！
 - ◆ Deep Open Intent Classification with Adaptive Decision Boundary
 - 论文: <https://arxiv.org/pdf/2012.10209.pdf>
 - 代码: <https://github.com/thuiar/Adaptive-Decision-Boundary>
 - ◆ Discovering New Intents with Deep Aligned Clustering
 - 论文: <https://arxiv.org/pdf/2012.08987.pdf>
 - 代码: <https://github.com/thuiar/DeepAligned-Clustering>
- ◎ 推荐内容
 - ◆ 开放知识发现阅读清单 (持续维护中)
 - 链接: <https://github.com/thuiar/Adaptive-Decision-Boundary>
 - 收集整理了NLP、CV、ML等领域开放场景知识发现的相关论文、代码及引用等资料





参考文献

- ◉ Walter J. Scheirer, Anderson de Rezende, Archana Sapkota and Terrance E. Boult . 2013. **Toward Open Set Recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- ◉ Geli Fei and Bing Liu. **Breaking the Closed World Assumption in Text Classification**. 2016. In Proceedings of HLT-NAACL 2016.
- ◉ Abhijit Bendale and Terrance E. Boult. 2016. **Towards Open Set Deep Networks**. In Proceedings of CVPR 2016.
- ◉ Lei Shu, Hu Xu and Bing Liu. **DOC: Deep Open Classification of Text Documents**. 2017. In Proceedings of EMNLP 2017.
- ◉ Ting-En Lin and Hua Xu. 2019. **Deep Unknown Intent Detection with Margin Loss**. In Proceedings of ACL 2019.
- ◉ Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu and Albert Y.S. Lam. 2020. **Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification**. In Proceedings of ACL 2020.
- ◉ Padmasundari and Srinivas Bangalore. 2018. **Intent Discovery Through Unsupervised Semantic Text Clustering**. In Proceedings of INTERSPEECH 2018.
- ◉ Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang and Lintao Zhang. 2018. **Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning**. In Proceedings of EMNLP.
- ◉ Dilek Hakkani-Tür, Yun-Cheng Ju, Geoff Zweig and Gokhan Tur. 2015. **Clustering Novel Intents in a Conversational Interaction System with Semantic Parsing**. In Proceedings of INTERSPEECH 2015.
- ◉ Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom and Zsolt Kira. 2019. **Multi-class classification without multi-class labels**. In Proceedings of ICLR 2019.





参考文献

- Yen-Chang Hsu and Zhaoyang Lv and Zsolt Kira. 2018. **Learning to Cluster in Order to Transfer Across Domains and Tasks**. In Proceedings of ICLR 2018.
- Ting-En Lin, Hua Xu and Hanlei Zhang. 2020. **Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement**. In Proceedings of AAAI 2020.
- Kai Han, Andrea Vedaldi and Andrew Zisserman. 2019. **Learning to Discover Novel Visual Categories via Deep Transfer Clustering**. In Proceedings of ICCV 2019.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. **Bert: Pre-training of deep bidirectional transformers for language understanding**. In Proceedings of NAACL 2019.
- Casanueva, I.; Temcinas, T.; Gerz, D.; Henderson, M.; and Vulic, I. 2020. **Efficient Intent Detection with Dual Sentence Encoders**. ACL WorkShop 2020.
- Larson Stefan, Mahendran Anish, Peper Joseph J, Clarke Christopher, Lee Andrew, Hill Parker, Kummerfeld Jonathan K, Leach Kevin, Laurenzano Michael A., Tang Lingjia and Mars Jason. 2019. **An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction**. In Proceedings of EMNLP-IJCNLP 2019.
- Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang and Hongwei Hao. 2015. **Short Text Clustering via Convolutional Neural Networks**. In Proceedings of ACL Workshop 2015.
- Caron Mathilde, Bojanowski Piotr, Joulin Armand and Douze Matthijs. 2018. **Deep Clustering for Unsupervised Learning of Visual Features**. In Proceedings of ECCV 2018.



谢谢大家观看！

如有任何问题请联系邮箱：zhang-hl20@mails.tsinghua.edu.cn

个人主页：<https://hanleizhang.github.io>





Graph-Enhanced Multi-Task Learning of Multi-Level Transition Dynamics for Session-based Recommendation

Presenter: Chao Huang

Outline

- ❑ Background
- ❑ Related Work
- ❑ The Proposed Framework
- ❑ Experimental Settings
- ❑ Evaluation Performance
- ❑ Conclusions and Future Work

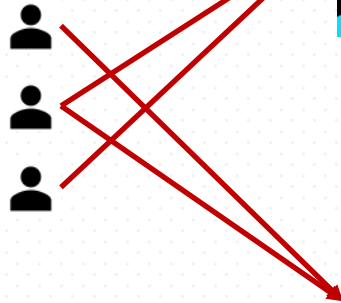
Recommender Systems

Recommender systems have play an important role in meeting user's personalized interests and alleviating the information overload for various applications

Online Advertising



E-commerce Platform



Collaborative Filtering

	HOTEL	T-shirt	Headphones	Smartphone
User 1	✓	✗	✓	✓
User 2	?	✓	✗	✗
User 3	✓	✓	✗	?
User 4	?	?	✓	✗

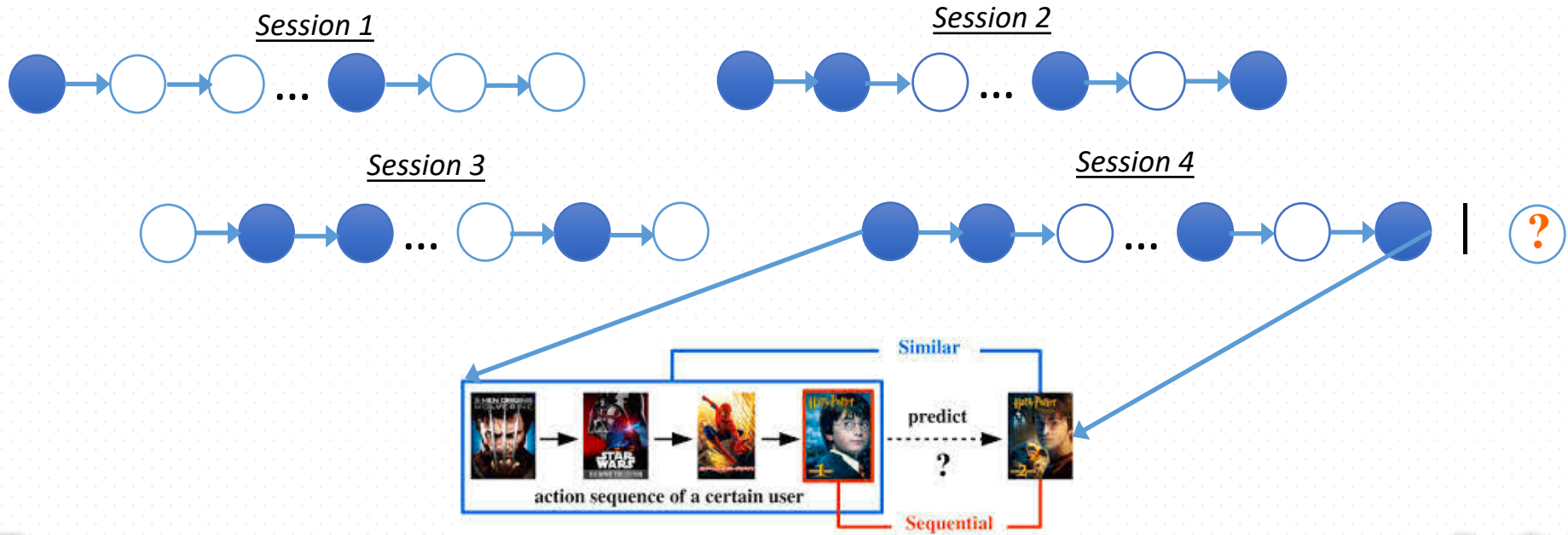
→

	HOTEL	T-shirt	Headphones	Smartphone
User 1	✓	✗	✓	✓
User 2	✓	✓	✗	✗
User 3	✓	✓	✗	✗
User 4	✓	✗	✓	✗

User-Item Interactions

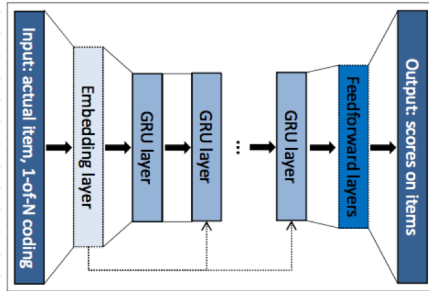
Session-based Recommendation

Session-based recommendation aims to predict the next interactive item based on a group of anonymous temporally-ordered behavior sequences of users



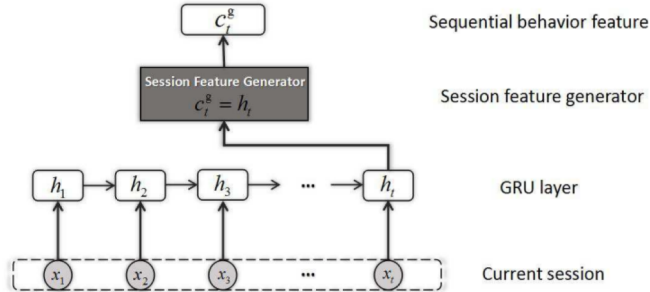
Related Work

Recurrent Session-based Recommender System



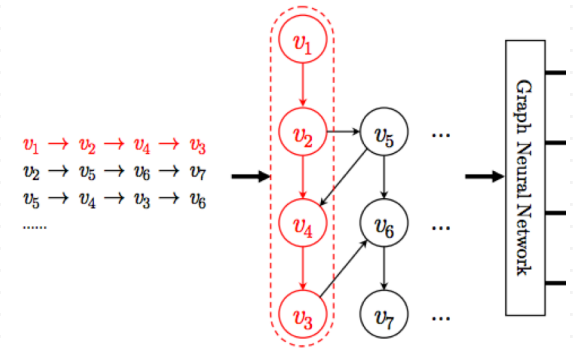
GRU4Rec

Attention-based Recommendation Frameworks



NARM

Session-based Recommendation with GNN



SR-GNN

Multi-Level Transition Dynamics

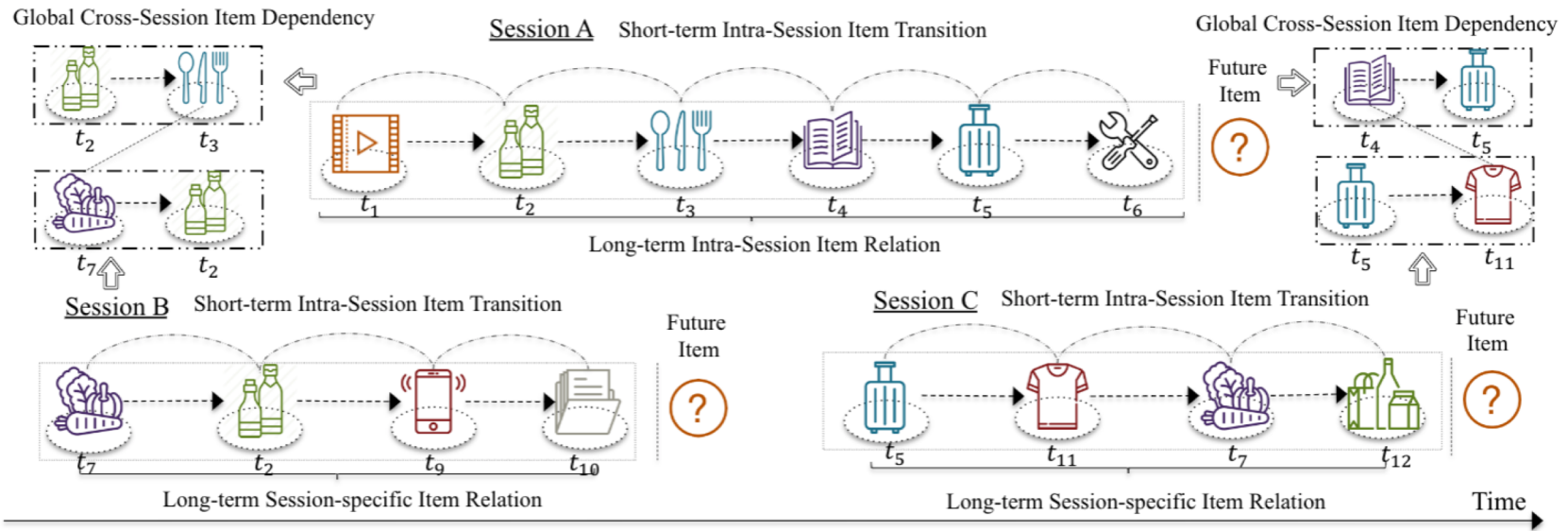
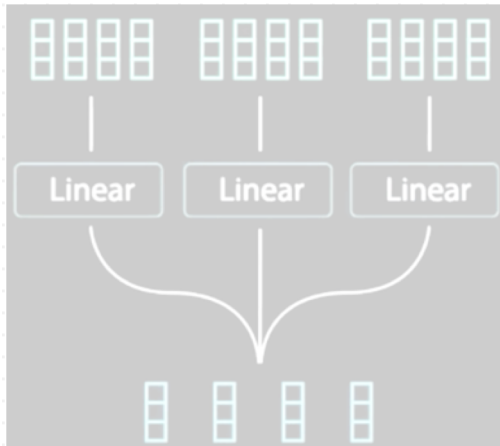


Fig. 1. Illustrated example of session-based recommendation with multi-level transition dynamics.

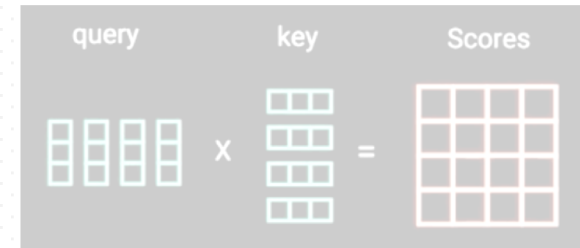
Intra-Session Item Relation Learning



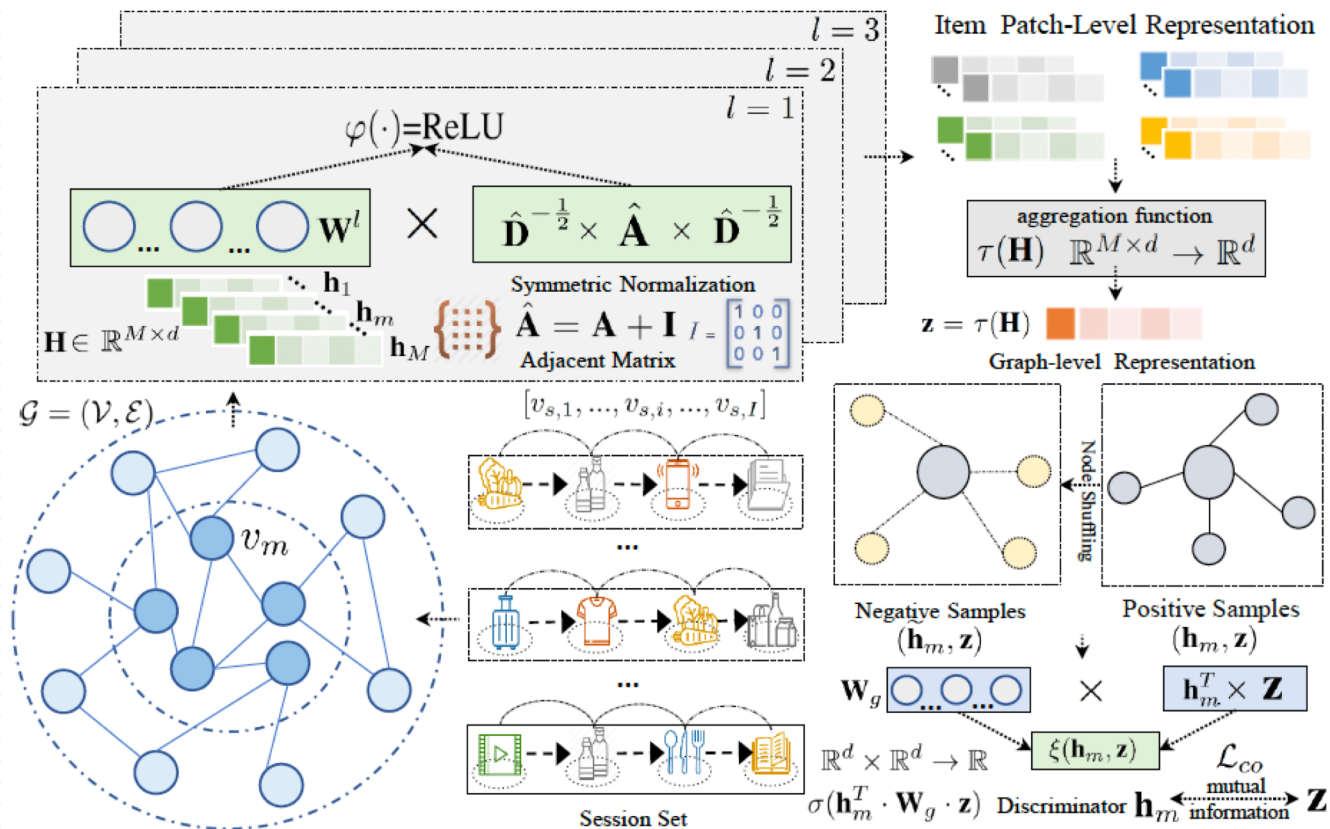
$$\begin{array}{l} X \times W^Q = Q \\ X \times W^K = K \\ X \times W^V = V \end{array}$$



Dot Product of Query and Key



Global Transition Dynamics Modeling



Experimental Settings

☆ Experimental Data and Evaluation Metrics

Dataset

Metrics

Table 1: Statistics of the experimented datasets.

Dataset	Yoochoose	Diginetica	RetailRocket
# Train Sessions	369,859	719,470	433,648
# Test Sessions	55,400	60,858	15,132
# All Items	17,376	43,097	36,968
Average Length	6.15	5.13	9.93

Precision@K (Pre@K)

Mean Reciprocal Rank@K (MRR@K)

☆ Compared Baselines

Frequency-based
Recommendation
Strategy

Neighboring
Relation Modeling
Algorithm

Recurrent Session-
based Recommender
System

Attention-based
Recommendation
Frameworks

Session-based
Recommendation
with GNN

Hybrid Session-based
Recommendation
Model

Performance Comparison

☆ Overall Performance

Table 2: Recommendation performance comparison of all methods in terms of Pre@10 and MRR@10

Dataset	Metric	POP	S-POP	It-KNN	GRURec	NARM	STAMP	SASRec	SR-GNN	CSRM	CoSAN	<i>MTD</i>
Diginetica	Pre	0.58	20.66	26.46	20.31	36.72	37.05	38.42	38.40	38.56	37.58	40.22
	MRR	0.19	13.59	10.91	7.78	15.00	16.05	16.27	17.04	16.23	15.57	17.58
Yoochoose	Pre	4.59	28.61	43.40	55.13	60.19	58.79	60.42	60.84	60.46	61.01	61.83
	MRR	1.51	18.45	21.39	25.76	29.03	29.44	30.47	30.57	30.37	30.21	30.83
Retailrocket	Pre	1.59	29.67	21.41	31.01	44.74	43.14	46.39	44.88	47.21	45.83	47.93
	MRR	0.44	21.51	9.78	15.37	25.54	26.65	26.74	26.95	27.14	26.01	28.51

☆ Model Ranking Performance

Table 3: Ranking performance with different K values.

Data	Metric	NARM	STAMP	SR-GNN	CSRM	CoSAN	<i>MTD</i>
Digi	Pre@5	24.80	25.72	27.15	26.38	25.72	28.29
	Pre@10	36.72	37.05	38.40	38.56	37.58	40.22
	Pre@20	50.32	49.86	51.57	52.56	50.94	53.92
Reta	Pre@5	36.25	36.45	37.38	38.65	37.07	39.64
	Pre@10	44.74	47.54	44.88	47.21	45.83	47.93
	Pre@20	52.58	55.56	52.27	55.04	54.87	55.95

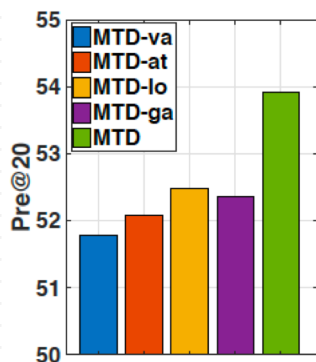
Model Ablation and Effect Analyses

☆ Effect of Hierarchical Attention Network

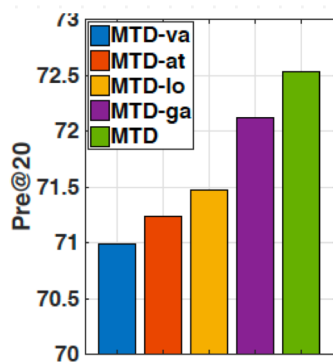
- i) MTD-va generates the session-level embeddings with the vanilla attention layer
- ii) MTD-at further incorporates the temporal factor into the MTD-va method

☆ Effect of Cross-Session Dependency Encoder

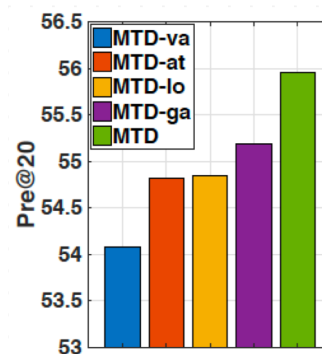
- i) MTD-lo only encodes the local-level item transition patterns without the cross-session dependency encoder
- ii) MTD-ga replaces our graph-structured hierarchical relation encoder with the graph attention network operated on all relevant sessions



(a) Diginetica

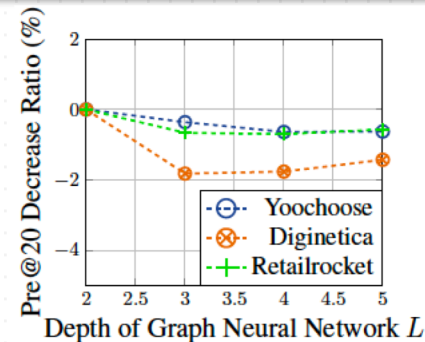
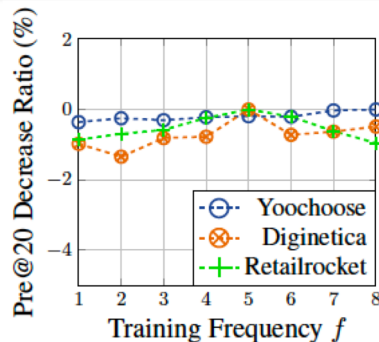
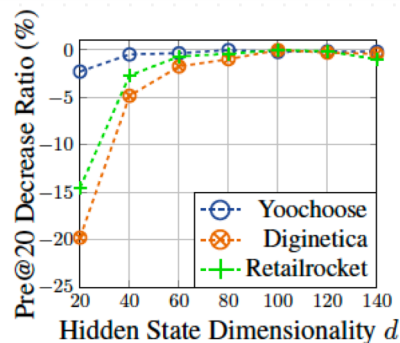


(b) Yoochoose



(c) Retailrocket

Hyperparameter Study



(1) Effect of Hidden Dimensionality d .

The performance saturates as the hidden dimensionality d reaches around 100.

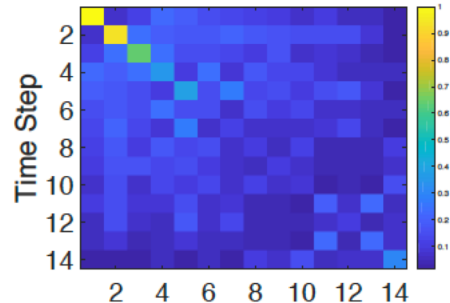
(2) Impact of Training Frequency f

A large value of f (≥ 5) will degrade the performance by misleading the objective function optimization

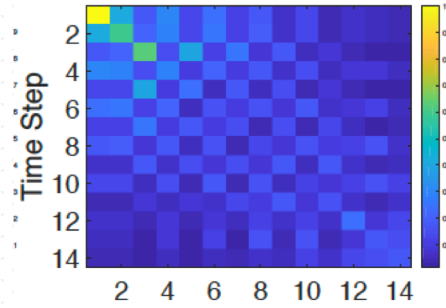
(3) Influence of Depth in Graph Neural Architecture

Stacking more graph convolution layers with the adjacent matrix-based aggregation will more involve more redundant information of high-order connectivity

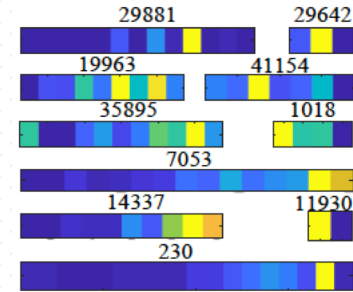
Case Study: Model Interpretation



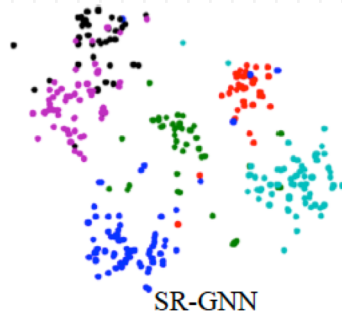
(a) Interpretation of pairwise item correlations of 1st case



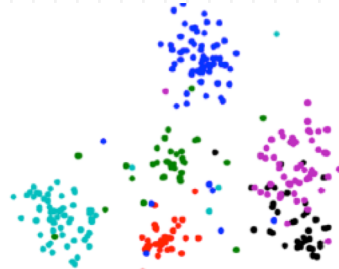
(b) Interpretation of pairwise item correlations of 2nd case



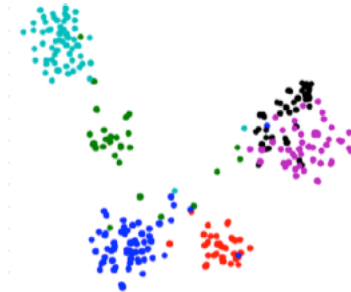
(c) Interpretation of item-specific importance within individual session



SR-GNN



STAMP



MTD

Conclusions and Future Work



This work develops a new multi-task learning framework–MTD



MTD not only models the intrasession sequential transitions, but also derives the high-order item relationships across long-range sessions



Experimental results on different real-world datasets show that MTD is superior to state-of-the-art baselines



In the future, we will incorporate item content information (e.g., item text description or reviews) into MTD to deal with external attributes in learning semantic-aware item transitions.

Acknowledgement



We are thankful to all the anonymous reviewers for their insightful feedback and suggestions



This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and the York Research Chairs (YRC) program.



This work is supported by National Nature Science Foundation of China (62072188, 61672241), Natural Science Foundation of Guangdong Province (2016A030308013)



Thank You



Invertible Concept-Based Explanations for CNN Models with Non-Negative Concept Activation Vectors

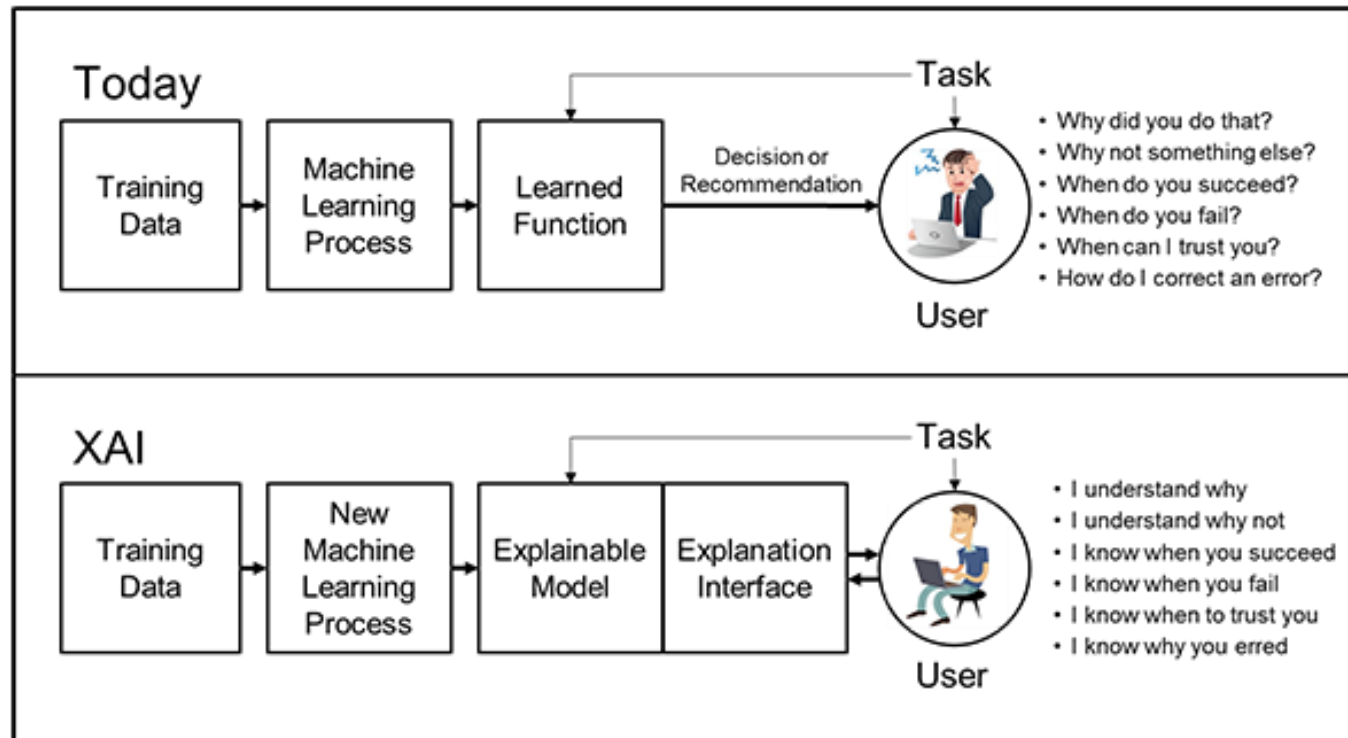
Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, Benjamin I. P. Rubinstein



What is explainability?

—

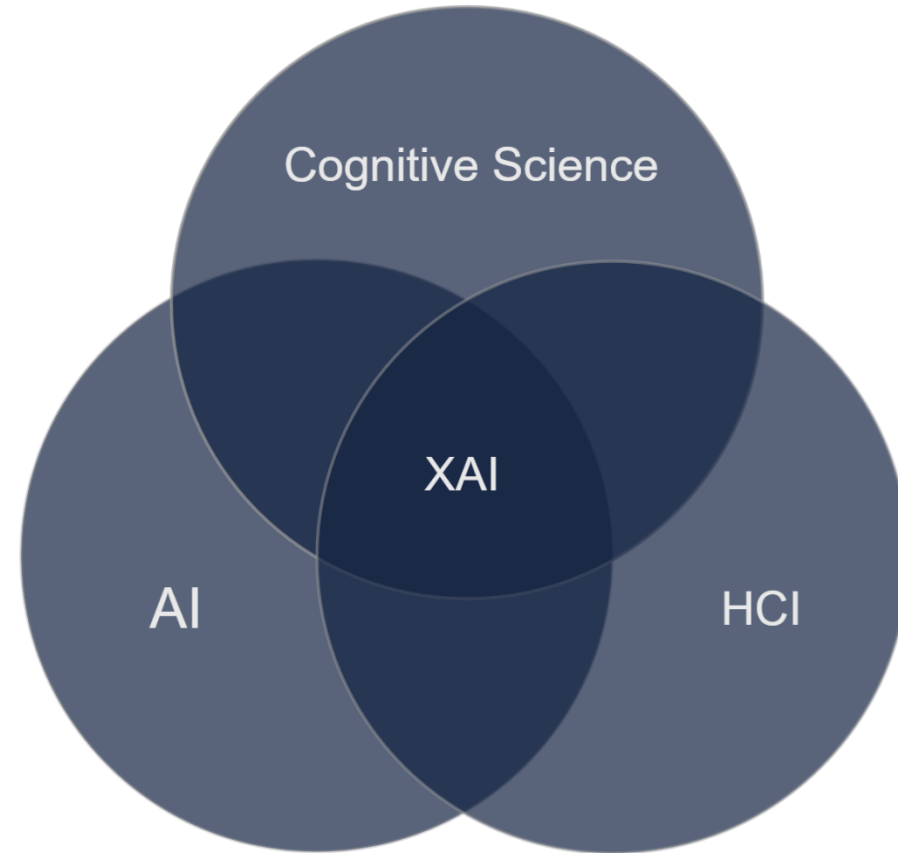
What is explainability?



eXplainable Artificial Intelligence (XAI):

- More transparent, interpretable, or explainable systems. To improve the performance.
- User will be better equipped to **understand** and therefore **trust** the intelligent agents

What is explainability?



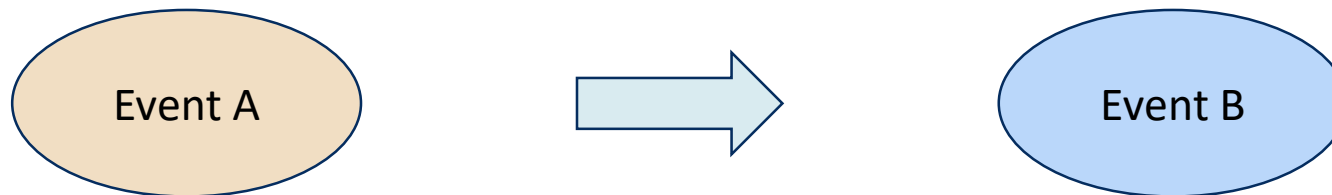


Explanations: Contrastive Explanation

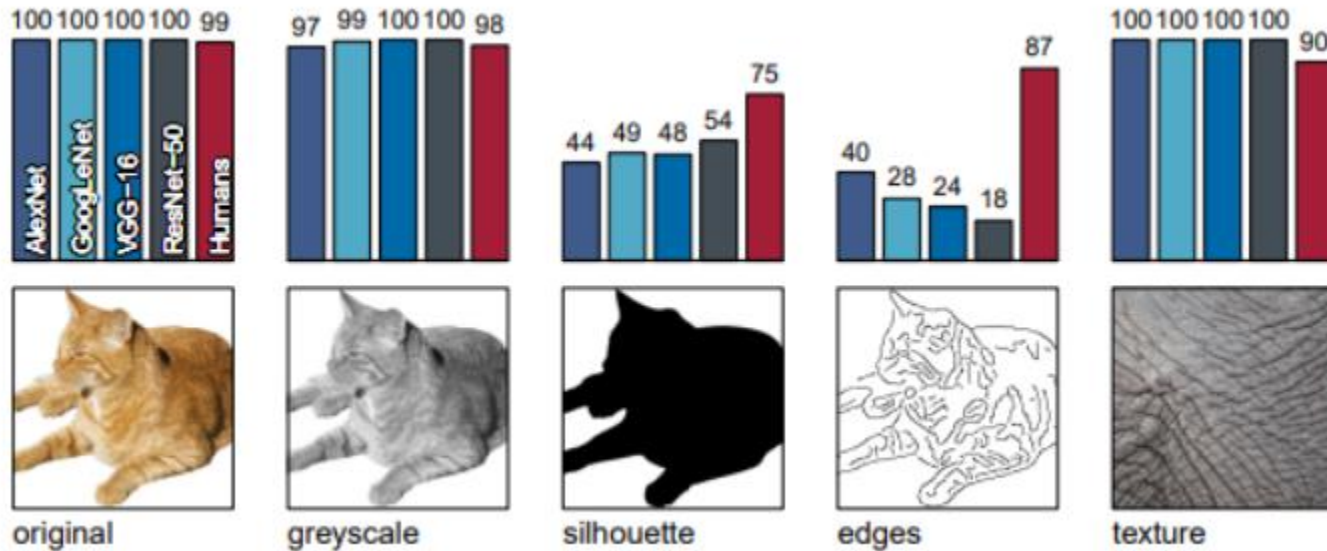
- *Why did Elizabeth open the door?*
 - *Why did Elizabeth open the door, rather than leave it closed?*
 - *Why did Elizabeth open the door rather than the window?*
 - *Why did Elizabeth open the door rather than Michael opening it*
- “Why P but not Q ?” for contrastive explanations.
 - P: facts, real, predictions
 - Q: foils, unreal, counterfactual case

Explanations: Causality

- Causal relationship: $A \gg B$
- Counterfactual: $\text{not } A \gg \text{not } B$
- Causal chains: $A_1 \gg A_2 \dots A_n \gg B$



What is explainability in CNN models?



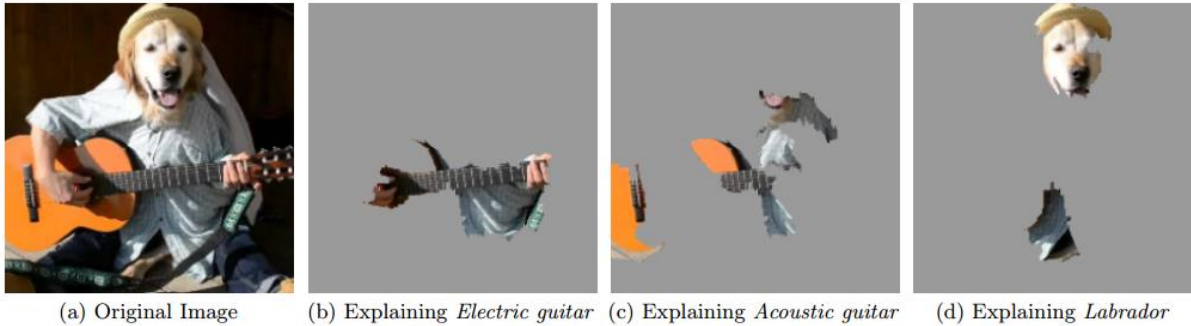
- CNN models predict images differently comparing with humans.
- Humans could predict silhouette and edges of cats correctly, but CNN model can not.



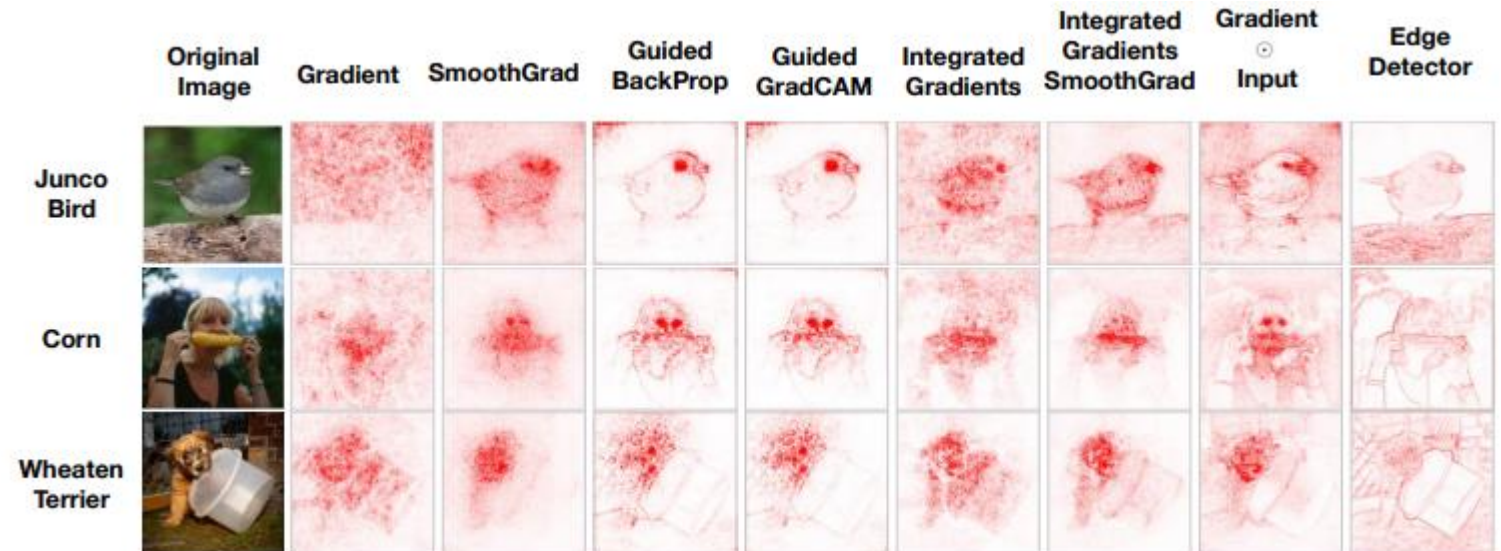
Related works

—

CNN explanations: Input level explanations



- LIME (upper left)
- CAM (lower left)
- Saliency maps (lower right)



CNN explanations: Interpretability inside

lamps in places net



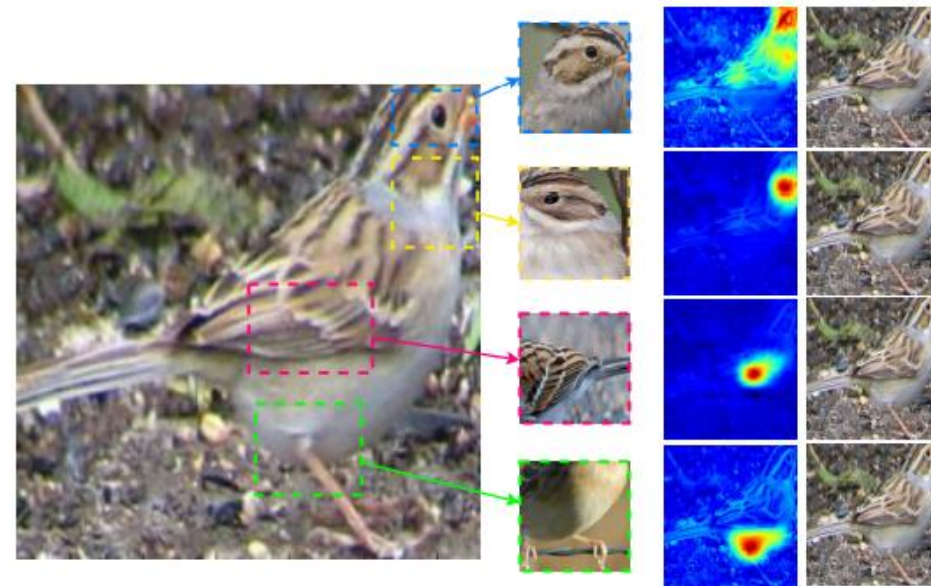
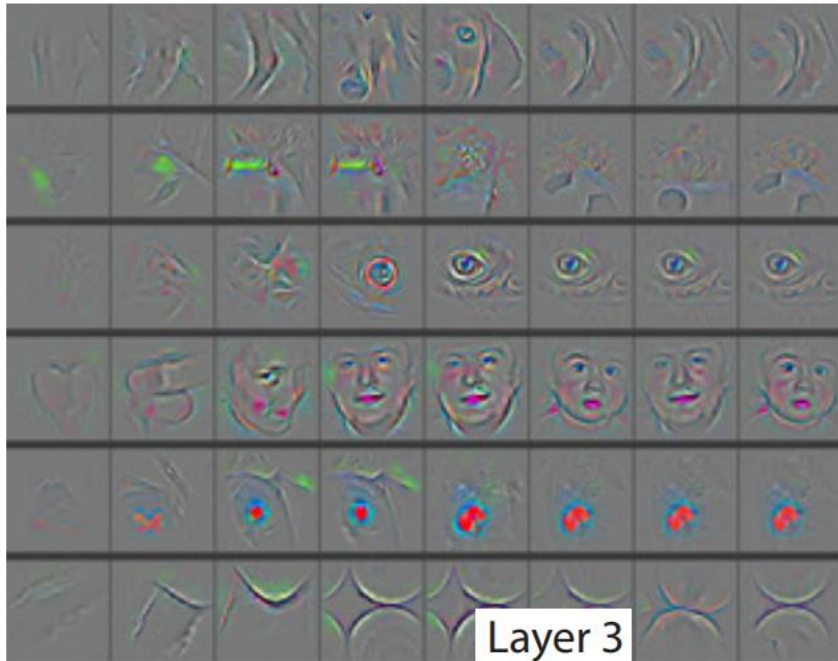
wheels in object net



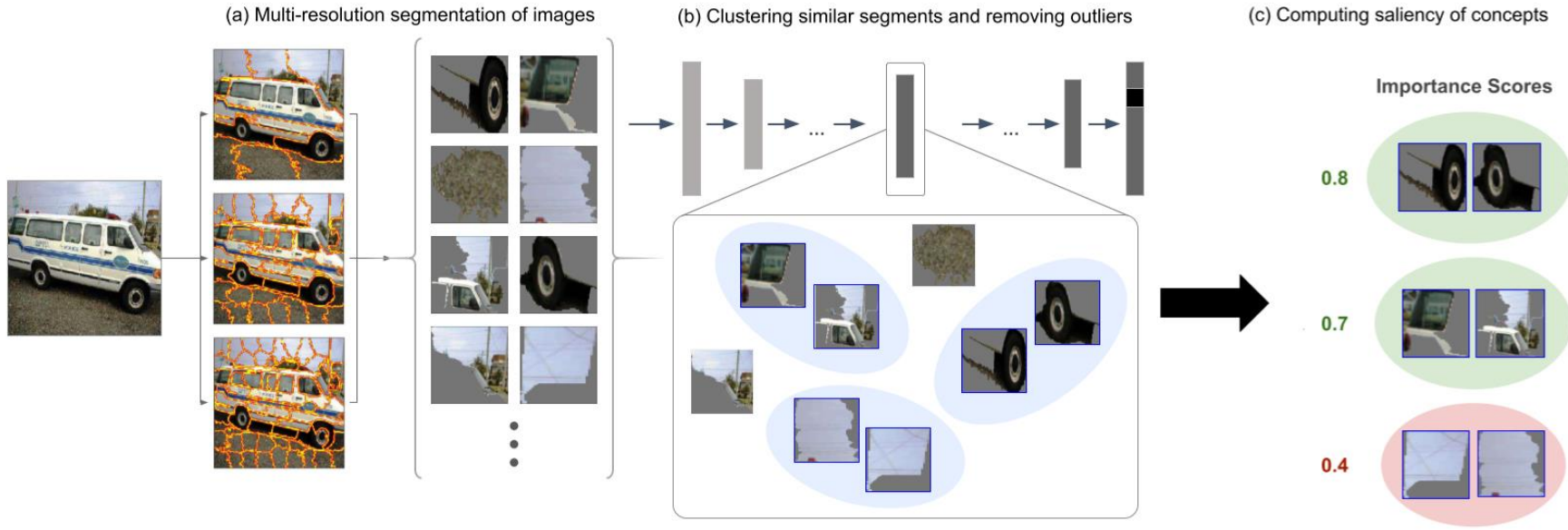
people in video net



- Network dissection (upper left)
- Deconvolution layer (lower left)
- Interpretable image recognition (lower right)



From TCAV and auto CAV



Concept activation vectors (CAV):

- Vectors inside the CNN model activate some concepts.
- Use ML models to get human understandable vectors from inside the model.
- TCAV and network dissection: linear model. Auto CAV: cluster method.



Invertible Concept-based Explanations

—

Concept-based Explanations

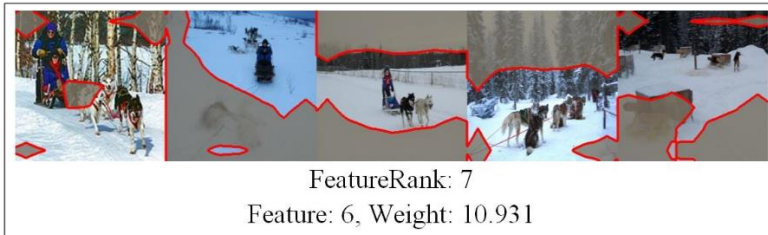
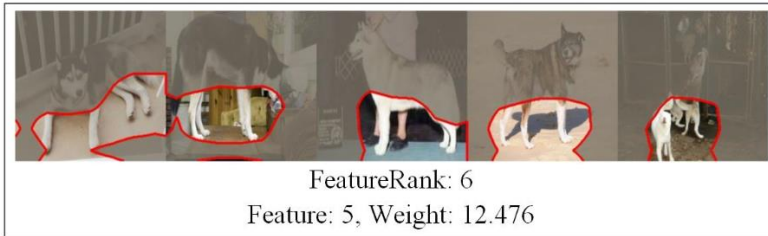
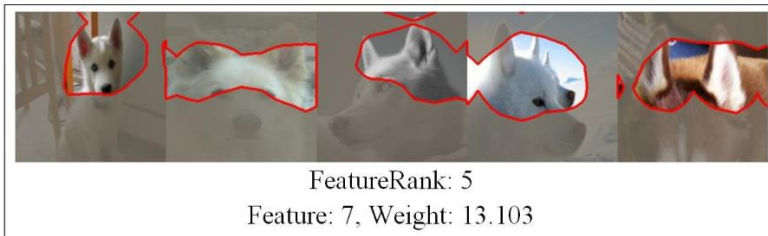


ClassName: tiger cat, Feature: 1

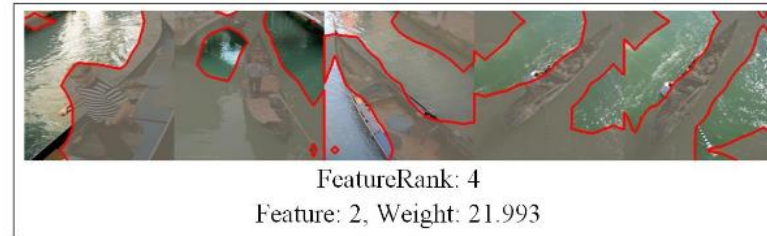
Similarity: 0.096, Weight: 24.416, Contribution: 2.336

Global Explanation

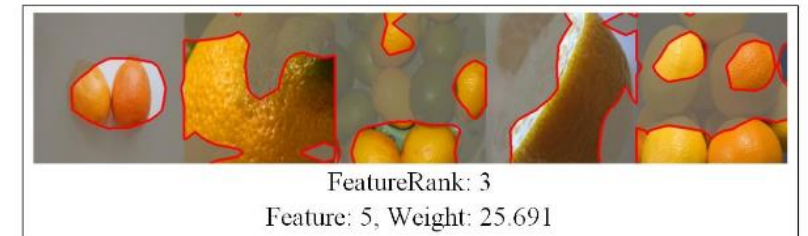
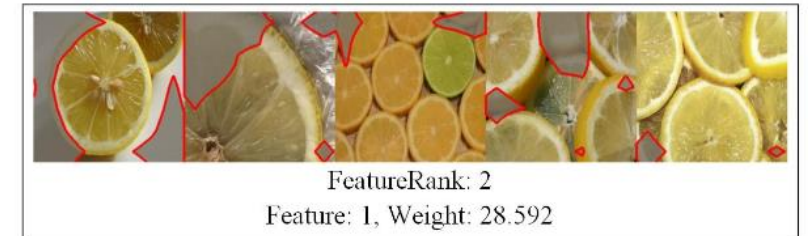
ClassName: Eskimo dog, husky
Fidelity error: 6.212 %



ClassName: gondola
Fidelity error: 5.620 %













ClassName: lemon
Fidelity error: 5.433 %



Local Explanation



Concept ID	Related Area	Concept Prototypes	Similarity Score	Concept Weight and Contributions for dog (up) and cat (down)	
0			0.091	$\times 37.031 = 3.359$ $\times 8.229 = 0.746$	<p>For dog: Concepts Score: 9.095 Residual Error: 0.291 CNN prediction: 9.386</p> <p>CNN: it's more like a dog!</p> <p>For cat: Concepts Score: 8.249 Residual Error: 0.311 CNN prediction: 8.560</p>
5			0.085	$\times 16.644 = 1.409$ $\times 0.496 = 0.042$	
...		...			
1			0.096	$\times -0.856 = -0.082$ $\times 24.416 = 2.336$	
6			0.140	$\times 4.102 = 0.573$ $\times 21.091 = 2.946$	
...		...			
2			0.117	$\times -1.911 = -0.223$ $\times -0.802 = -0.094$	



Framework of the unsupervised explainer

Assumptions of CNN explanations:

- Using a linear model to approximate the CNN model F :
 - Linear models are simple and explainable
 - Consider weights w as importance and explanations

$$y = F(x) \rightarrow y' = wx' + b$$

- x' are image segments in LIME, pixels in saliency maps and image regions in CAM
- x' are some concepts in concept-based explanations

Framework of the unsupervised explainer

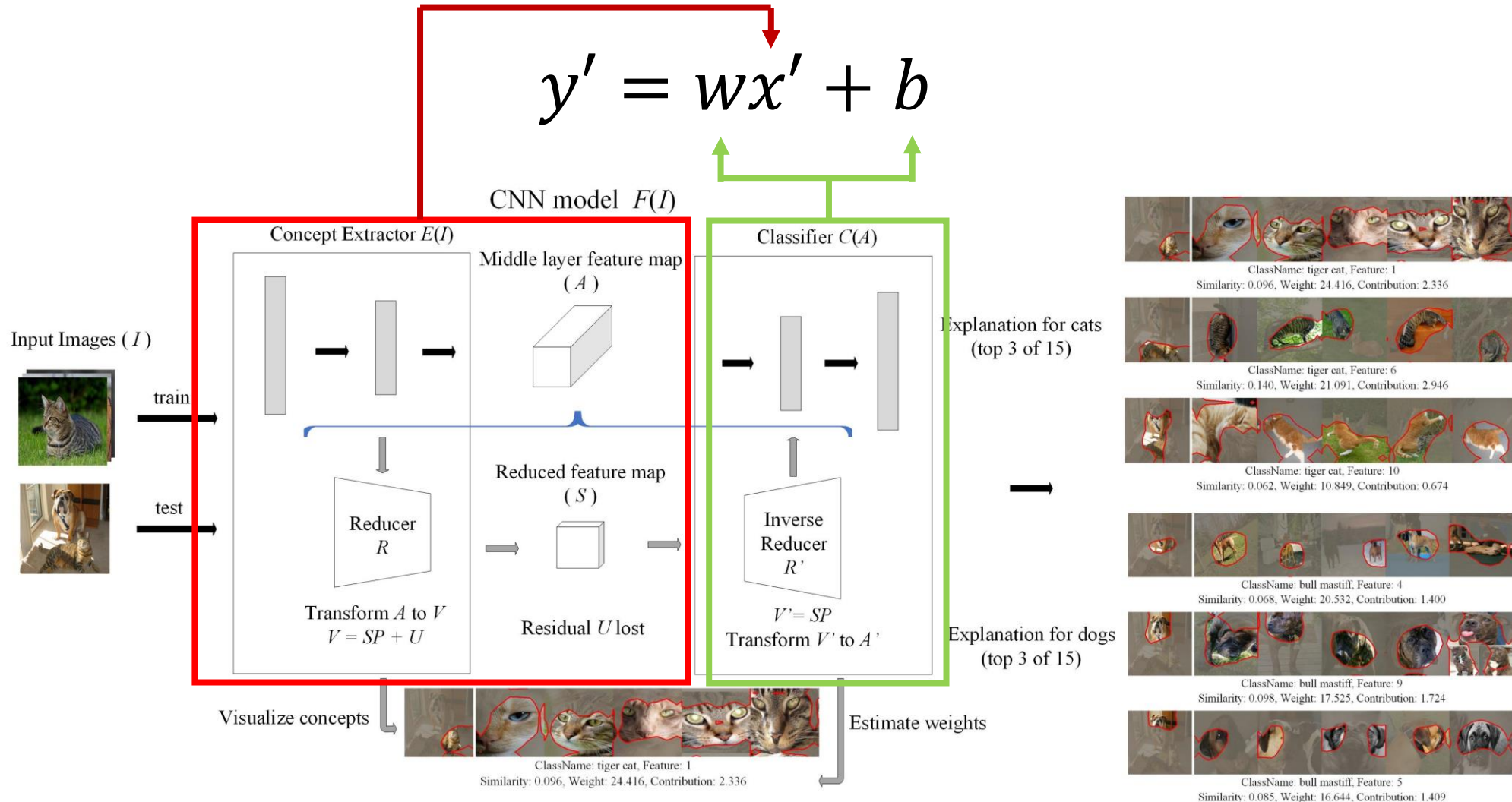
$$y = F(x) \rightarrow y' = wx' + b$$

Our target:

- x' should be meaningful to human
- w should be accurate to the CNN model F
- y' should be close to the prediction y
- Number of features should be acceptable

Framework of the unsupervised explainer

Separate the CNN model





Framework of the unsupervised explainer

Separate the CNN model

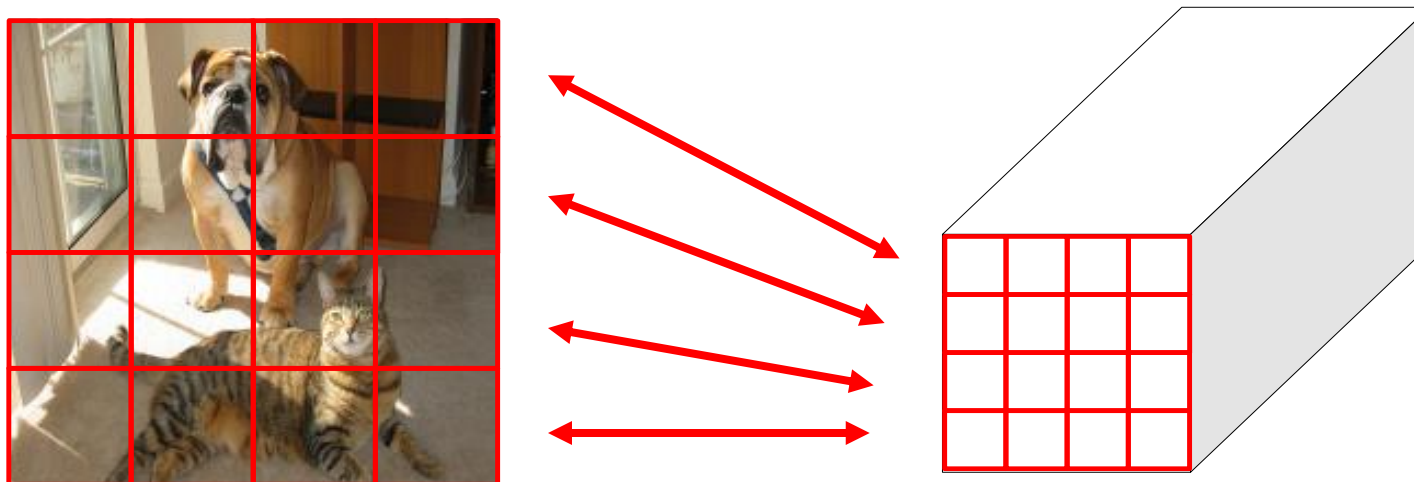
$$y' = wx' + b$$

- The whole CNN model is separated into feature extractor and classifier.
 - Higher layers includes higher level of features.
 - Simpler classifiers are easier to approximate → more accurate w .
- → Last layer is the best choice for separation.
 - w and b will be stable if only a GAP and dense layer at last.

Framework of the unsupervised explainer

Reduce the number of channels through matrix factorization:

- Consider each position in feature map as an instance and reduce the channel dimension. It's trained over an image set with some concepts.
- Points in feature maps are correlated to certain areas in the images. It disentangles the concepts in the feature maps.



Framework of the unsupervised explainer

Visualize the features

- Select images with the highest activation values on the target feature.
- Use CAM to generate heatmap and visualize the feature on each image.



ClassName: tiger cat, Feature: 1

Similarity: 0.096, Weight: 24.416, Contribution: 2.336



Evaluation

—



Evaluation

$$y = F(x) \rightarrow y' = wx' + b$$

Our target:

- x' should be meaningful to human (**Interpretability**)
- w should be accurate to the CNN model F (Done)
- y' should be close to the prediction y (**Fidelity**)
- Number of features should be acceptable (Done)



Evaluation

Experiment:

- Factorization methods (reducer):
 - PCA, NMF and K-means (used in ACE)
- CNN models and datasets:
 - ILSVRC2012: ResNet50 and Inception-V3 from TorchVIsion
 - Layer4 for ResNet50 and mixed_7c for Inception-V3 as target layer
 - CUB: ResNet50 with top1 error 15.81%
 - Layer4 as target layer



Evaluation

Fidelity measurement:

- Measure performance of the linear approximate model over different c' (feature number)
- Evaluate on NMF, PCA and K-means (Cluster) on test set.

- For classification model: Fid_c

- For regression model: Fid_r

$$Fid_{c_{F,\hat{F}}}(I) = \frac{\#\{i \in I \mid F(i) = \hat{F}(i)\}}{\#\{I\}}$$

$$Fid_{r_{F,\hat{F}}}(I) = \frac{\sum_{i \in I} |F(i) - \hat{F}(i)|}{\sum_{i \in I} |F(i)| + \epsilon}$$

Evaluation

Interpretability measurement:

- Task prediction: people can predict the explainer's prediction if explainer is understood.

Participants are required to:

- Classify the feature in the given image
 - Measure classification accuracy
- Name all the features
 - Measure pair-wise description distances (GloVe)
- Score the confidence to the answer
- Score the explanation quality



Evaluation

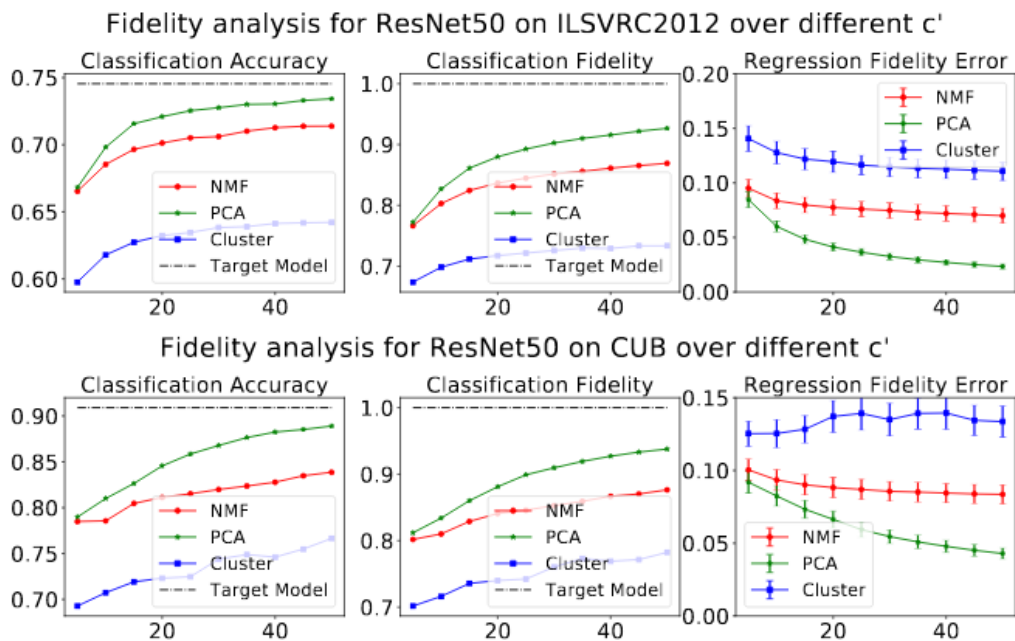


Table 1: Top: Mean and standard deviation of prediction accuracy, description similarity, confidence and quality comparison for 9 different groups. Middle: ANOVA test p values for each scenario. Bottom: T-test p values for each pair of reducers

Scenario	Reducer type	Accuracy	Description Similarity	Confidence	Quality			
					Understand	Satisfaction	Sufficiency	Completeness
RI	NMF	74.4% ± 9.2%	0.59 ± 0.1	77.7% ± 13.0%	4.3 ± 0.6	4.1 ± 0.6	3.8 ± 0.8	3.7 ± 1.2
	Cluster	66.3% ± 13.8%	0.56 ± 0.08	75.6% ± 13.8%	4.2 ± 0.7	3.8 ± 1.0	3.9 ± 1.0	3.6 ± 1.1
	PCA	37.8% ± 5.9%	0.52 ± 0.08	78.3% ± 14.7%	4.0 ± 0.9	3.8 ± 1.1	3.8 ± 1.1	3.7 ± 1.2
II	NMF	62.6% ± 18.6%	0.57 ± 0.08	69.3% ± 13.2%	3.5 ± 1.0	3.4 ± 1.3	3.3 ± 1.1	3.4 ± 1.3
	Cluster	44.8% ± 13.2%	0.53 ± 0.09	75.1% ± 13.7%	3.9 ± 1.1	3.6 ± 1.2	3.6 ± 1.2	3.5 ± 1.4
	PCA	40.0% ± 8.6%	0.49 ± 0.08	76.0% ± 13.0%	3.8 ± 0.9	3.7 ± 1.1	3.4 ± 1.2	3.2 ± 1.3
RC	NMF	81.1% ± 8.4%	0.7 ± 0.04	79.5% ± 10.8%	4.1 ± 0.8	3.7 ± 0.9	3.4 ± 1.2	3.5 ± 1.1
	Cluster	78.6% ± 15.5%	0.7 ± 0.05	75.0% ± 18.7%	3.9 ± 1.0	4.1 ± 1.0	4.0 ± 1.0	3.9 ± 1.1
	PCA	57.0% ± 11.6%	0.59 ± 0.03	61.1% ± 17.6%	3.6 ± 1.0	3.0 ± 1.2	3.4 ± 1.2	3.2 ± 1.2

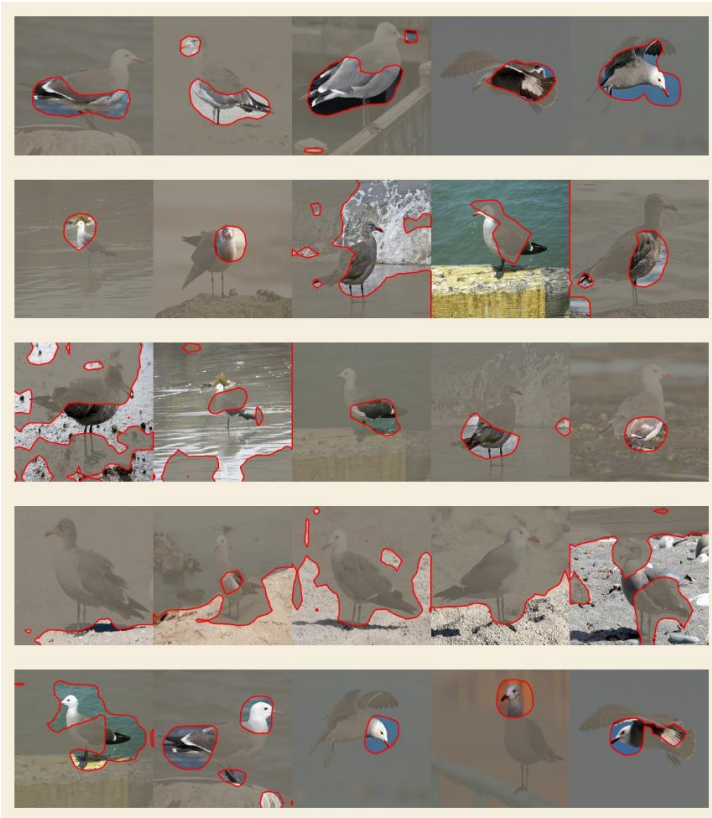
Scenario	Accuracy	Description Similarity	Confidence	Quality			
				Understand	Satisfaction	Sufficiency	Completeness
RI	<0.001	0.131	0.841	0.446	0.592	0.941	0.948
II	<0.001	0.064	0.304	0.493	0.752	0.690	0.844
RC	<0.001	<0.001	0.004	0.283	0.016	0.219	0.174

Scenario	Reducer Pair	Accuracy	Description Similarity	Confidence	Quality			
					Understand	Satisfaction	Sufficiency	Completeness
RI	NMF + Cluster	0.053	0.454	0.650	0.489	0.328	0.743	0.744
	NMF + PCA	<0.001	0.058	0.904	0.222	0.350	1.00	0.893
	Cluster + PCA	<0.001	0.182	0.589	0.549	0.979	0.783	0.849
II	NMF + Cluster	0.006	0.298	0.246	0.277	0.693	0.387	0.911
	NMF + PCA	<0.001	0.016	0.152	0.429	0.451	0.739	0.643
	Cluster + PCA	0.251	0.204	0.864	0.659	0.762	0.604	0.597
RC	NMF + Cluster	0.558	0.863	0.402	0.605	0.261	0.143	0.293
	NMF + PCA	<0.001	<0.001	<0.001	0.116	0.074	0.894	0.411
	Cluster + PCA	<0.001	<0.001	0.029	0.304	0.008	0.108	0.068

Evaluation

Survey Samples:

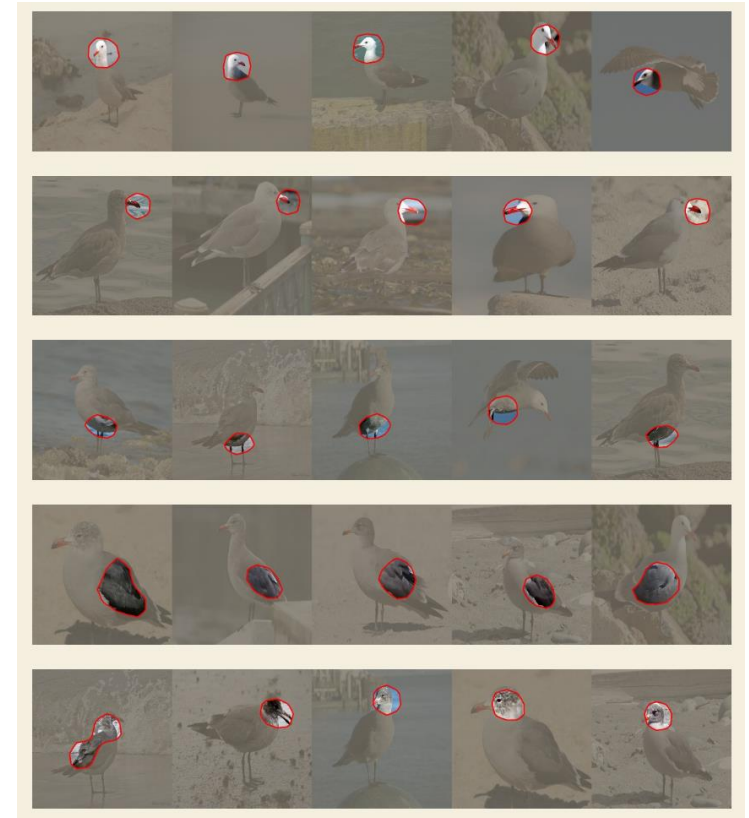
PCA



K-Means



NMF





Conclusion

—

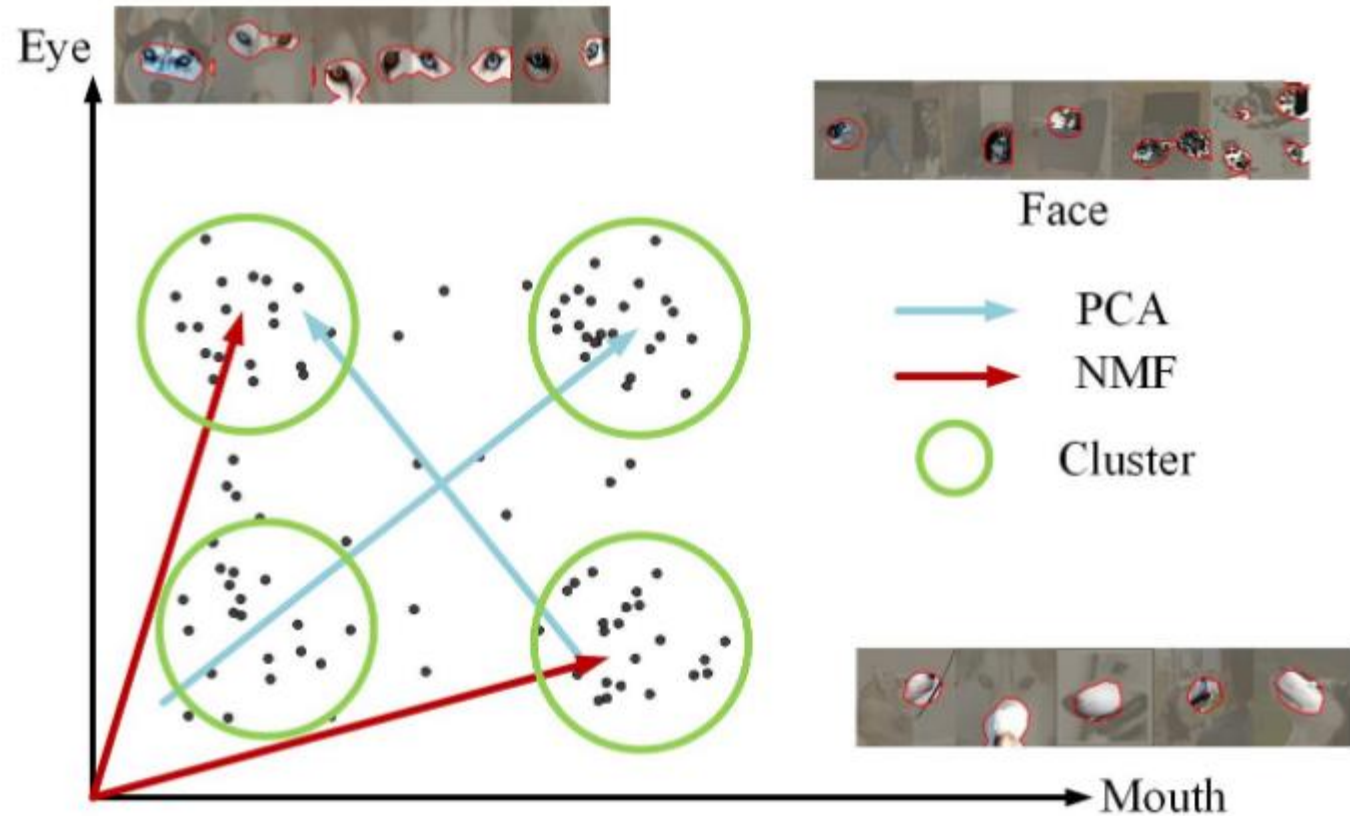


Conclusion

- Use linear approximate models with matrix factorization reducer
- For fidelity:
 - PCA > NMF >>> K-means
 - Explainers with NMF or PCA can replace the original model with around 80% to 90% performance.
- For interpretability:
 - NMF > K-means >>> PCA
- For explanations, interpretability > fidelity
- → NMF provide the best explanations and NCAVs (Non-negative CAVs) with a little worse fidelity.

Conclusion

- PCA
 - Wrong CAVs
- K-means
 - Repeat CAVs
- NMF:
 - The most accurate





THE UNIVERSITY OF
MELBOURNE

Thank you

