

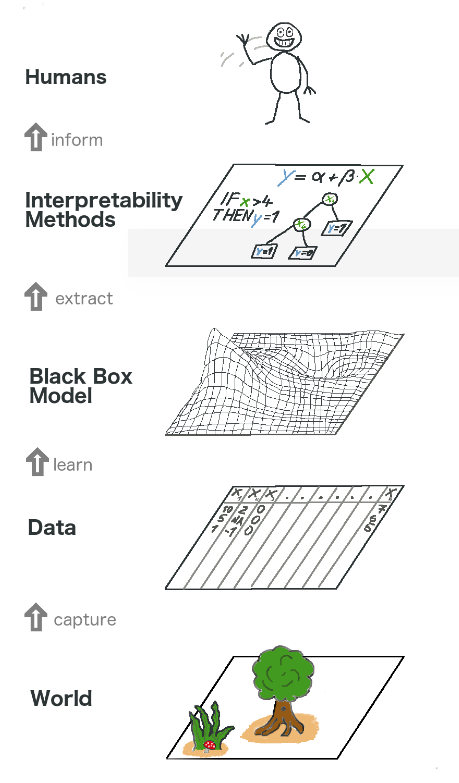
KDD 2020

目标：高效的模型无关的实例特征选择（IFS）方法，其目标在于解决现有IFS方法中存在的完备性（sanity）、组合捷径（combinatorial shortcuts）、模型可识别性（model identifiability）和信息传递（information transmission）四个方面的问题。

方式：将原模型输出作为解释器的额外输入，增加针对于未选中特征的对抗学习机制以辅助学习选中特征与目标的条件概率，将其他高效的解释性方法作为先验实现热启动。该方法在文本数据和图像数据以及时间序列数据的五个基准数据集中的四个都达到了最佳性能，在没有达到最佳精度的数据集上也表现出接近最佳的效果，并且有着最好的鲁棒性。

**什么是模型可解释性**

模型可解释性表达了模型内在机制的透明度以及人类理解模型决策原因的难易程度，主要体现在两个方面：

* **为什么模型会做出某种决策？**对于一个分类任务，当往模型中输入一个样本时会得到一个预测，模型的可解释性帮助我们去确定模型为什么会产生这一预测。更具体地，样本的哪些特征使模型做出了这一预测。
* **人类能否理解并信任这一决策？**对于任何希望模型能预期工作的人而言，模型的解释必须是易于理解的，否则无法轻易信任模型或是对模型进行针对性的调整。

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “**Model-agnostic interpretability of machine learning.**” ICML Workshop on Human Interpretability in Machine Learning. (2016).

**为什么模型需要可解释性**

在现实场景中，模型可解释性和模型性能之间往往需要权衡。机器学习模型、深度学习模型的产生是为了解决复杂的显示问题，模型可以学习数据中的潜在模式和关系，这种模式不一定是人类所能理解的。在很多高风险领域如金融、医疗、犯罪领域，最终往往需要使用传统的机器学习模型，因为模型的可解释性对于工作者的实际决策密切相关。模型需要可解释性通常考虑到以下三个原因：

* **模型改进**：理解指标特征、分类、预测，进而理解为什么一个机器学习模型会做出这样的决定、什么特征在决定中起最重要作用，能让我们判断模型是否符合常理。假设一个场景：使用一个深度神经网络来学习区分狼和哈士奇的图像，并获得了90%+的准确率。但仅通过准确率我们无法判断是否模型仅依赖雪地背景就做出判断。如果模型构建者和使用者能知道模型是如何使用特征进行预测的，就能通过直觉判断模型是否使用了有意义的特征，模型是或否能泛化到其他样本的预测上。
* **模型可信性与透明度：**让黑箱模型来决定人们的生活是不现实的，尤其金融、医疗、犯罪领域。模型做出错误决策的代价往往是难以接受的。模型在医疗预测任务中可能非常准确，但是依然需要专家对诊断结果进行解释。可解释性有助于使用者信任和模型来支持他们工作。长久来看，更好地理解机器学习模型可以节省大量时间、防止收入损失。如果一个模型没有做出合理的决定，在应用这个模型并造成不良影响之前就可以发现这一点。
* **识别或防止偏差：**有偏差的模型经常由有偏见的事实导致，如果数据包含微妙的偏差，模型就会学习下来并认为拟合很好。一个例子是犯罪量刑领域，模型可能会学习到包含种族偏见的模式，可解释性可以帮助人们预防偏差的增大甚至是消除不公平的偏差。

**可解释性分类**

* **Model-specific：**解释方法依赖于模型的参数和使用的特征,即这个模型是可以产生解释的.例如:回归模型的系数,决策树的规则.
* **Model-agnostic：**解释方法是一种事后的方法,通常通过分析输入和输出对来操作,这些方法无法访问任何模型的内部. (MEED=Model-agnostic ..)

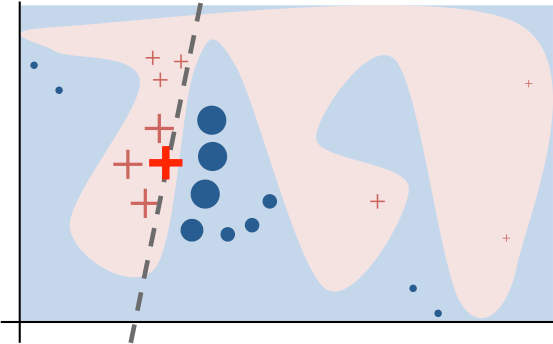
本文是基于实例的特征选择(IFS),IFS为每个样本生成一个特征重要度的分数，这分数表明了对于某一样本而言哪些特征对产生对应的输出起着至关重要的作用。和attention很相似,但是又有所区别.

(对抗)注意力机制: 消除在不同学习任务的情况下,特征提取的差异.

特征选择: 随着输入的不同,导致输出不同.将标签预测器的预测作为另一个输入信号处理,以便了解关于正确标签的信息.

**当前工作**

* **Feature attribution methods：**特征归因方法可以分为全局方法和局部方法。这类方法通过扰动特征来探究特征的对模型输出的影响。例如LIME方法



* **Direct model-interpretation (DMI) methods：**tree-based and rule-based两种.都是通过模型直接解释.但是,如果原始模型极为复杂,那么这个方法通常是无效的.例如L2X和VIBI方法,去学习拟合器基于神经网络,即他们学习的是特征,而不是输出,上述4个问题都有可能发生.

**可解释性需要考虑的问题**

* Expressiveness：这一属性指出能获得高分数的特征的数量应当是较少的。一个直接的理解应该是重要特征和不重要特征间区分度应该较大。
* Fidelity：保真度这一属性指出模型的输出应当主要由高分的特征所决定。
* Low sensitivity：低敏感这一属性指出生成的特征分数应该是高鲁棒性的，对对抗样本的攻击是不敏感的。
* Sanity：这一属性指出生成的特征分数应当取决于被解释的模型。需要注意的是，前面提到的LIME方法中获得的特征分数更多是针对于用于解释的模型而不是针对于需要解释的模型。

**问题描述**

考虑一个数据集包含有个独立的样本，其中第个样本记为，数据驱动的黑盒模型，模型输出 。IFS问题需要构建一个解释器，它的输出是一个特征重要性得分向量。换言之，解释器需要建立起一个映射，但由于黑盒模型无法直接作为神经网络的输入，因此通常使用替代映射。

**本文解决的问题**

**Sanity problem**

解释器选中的特征可能是与原模型无关，而仅仅只与输入的样本有关。这意味着选中的特征可能和原模型在预测中真正使用的特征是不一致的。这要求生成的解释具有Sanity这一属性。

**Combinatorial shortcuts problem**

解释器选中的特征可能并不是良好的特征，解释模型可能将生成的mask作为额外的特征以辅助数据和标签的拟合。举个例子，解释模型可以对每个样本都选择取前半部分或后半部分进行拟合，解释模型将会关注这种模式是否对性能有提升，而不是考虑是否是因为选择了好的特征才使得性能提升。这要求生成的解释需要具有Fidelity(忠诚度)这一属性。

**Model identifiability problem**

解释器可能会产生多种具有相似性能的特征组合，解释器很难确定哪一种组合才是最好的。这要求生成的解释需要具有可表达性这一属性。

**Information transmission problem**

解释器生成特征得分向量的过程是无监督的，因此难以将监督信息传递给解释器，解释器也很难利用好监督信息，因此解释器训练起来难度很大。

**本文解决方法**

**将原模型输出作为解释器的额外信号(解决问题1和4)**

现有的很多方法直接将原样本输入到解释器中，这一过程没有黑盒模型的参与，这往往会产生Sanity Problem。因此该方法将原模型的输出也作为解释器的一个输入，可以加强生成的特征得分向量与原模型间的联系。另外这一策略为解释器提供了额外的信息，这可以解释器能学习到更多的知识，在一定程度上也能减轻Information transmission problem。

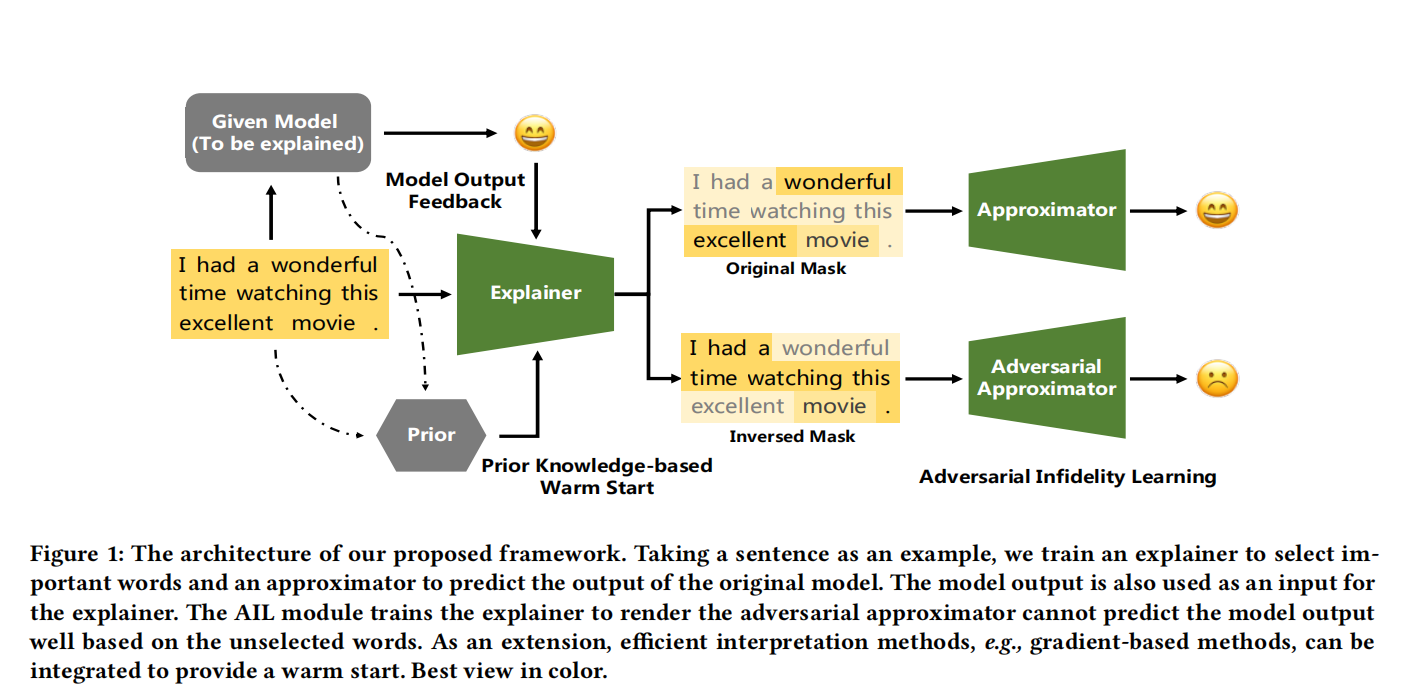
**针对于未选中特征的对抗学习机制（AIL）(解决问题2和3)**

AIL机制的提出是为了解决combinatorial shortcuts problem和model identifiability problem。简而言之，希望解释器选中的特征组合是足够好且唯一的，而未选中的特征包含的都应该是无用的信息。基于此想法，AIL机制中增加了一个拟合器（Approximator），使用它来拟合未选中特征和模型的输出，解释器的目标是使这个Approximator的精度尽可能小。

**基于先验知识的暖启动(解决问题4)**

解释器的训练本身存在有Information transmission problem，再加入AIL对抗学习机制后由于对抗学习的不稳定性导致模型更难以收敛。为此，论文提出集成其他高效模型的解释和先验用作MEED模型的暖启动，在训练进行到一定程度，可以学到更好的解释器后，先验的约束就会逐渐放宽，类似于知识库的想法。

**MEED结构**



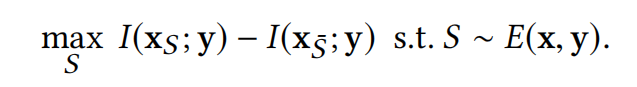
基本上，MEED的一个核心就是在于GAN的对抗模式与原始模型的输入。

拟合器：拟合黑盒模型，不断依次迭代拟合器和解释器，从而达到最优解。

**AIL机制**

**互信息**

互信息(Mutual Information)是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。公式：I(X;Y)=H(Y)−H(Y|X)。最大值：X=Y，即此时两者信息相等，包含的信息量最多。

**选什么特征？**  


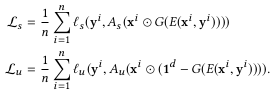
选中的特征对y的互信息要越大越好，即这些特征应当包括y。未选中的越少越好，这些特征这些特征和Y的差别很大。

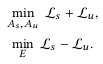
**损失函数**

待解决问题（变分下界）：



转换后的损失函数：（Ls,Lu就是逼近器As,Au的逼近（拟合）损失）





Step 1. 固定解释器E，对As和Au进行训练，这一过程使Ls和Lu都尽可能小，这意味着两个逼近器会被拟合得很好。

Step 2. 固定As和Au，对解释器E进行训练，这一过程会破坏Au的精度使Lu增大以达到优化目标。

这两个过程交替迭代，解释器E和逼近器Au的训练呈现出对抗的局面，这迫使解释器找到一种划分方式使Au无论训练都无法很好地逼近。可以理解为解释器E找到了使未选中特征包含最少的有用信息的划分方式，进而得到了高质量的选中特征。

**实验**

**Model-agnostic baselines：**LIME , kernel SHAP , CXPlain(CXP) , INFD , L2X , VIBI

**Model-specific baselines：**Gradient (Grad) , Gradient×Input (GI)

忠诚度指标

FS-M和FS-A应该尽可能高，说明黑盒模型的输出依赖于选中的特征。

FU-M和FU-A应该尽可能低，说明没有选中的特征对黑盒模型的影响很小。

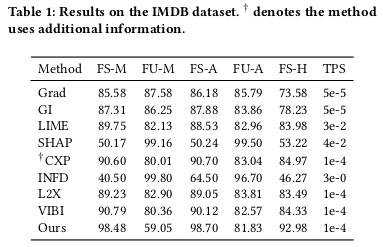
需要注意的是，如果选中特征的数量很少，可能会使A′s拟合效果不好而Au′拟合得很好，表现为FS-A较低以及FU-A较高。

以上提到的四个指标用于验证选中的特征能否很好解释黑盒模型是怎样产生预测的，在此之上可解释性还要求模型产生的解释尽可能让人容易理解，因此引入了FS-H指标。

| **指标** | **二者保真度/含义** |
| --- | --- |
| FS-M (%) | M(x)和M(被选用作解释的特征) |
| FS-A (%) | M(x)和A(被选用作解释的特征) |
| FU-M (%) | M(-x)和M(未被选用作解释的特征) |
| FU-A (%) | M(-x)和A(未被选用作解释的特征) |
| FS-H (%) | M(x)和人类使用”被选用作解释的特征“产生的判断 |
| SEN (%) | 对抗样本对特征分数的影响 |
| TPS | 每个样本获取预测的平均用时 |

**实验1/3：文本数据（IMDB）**

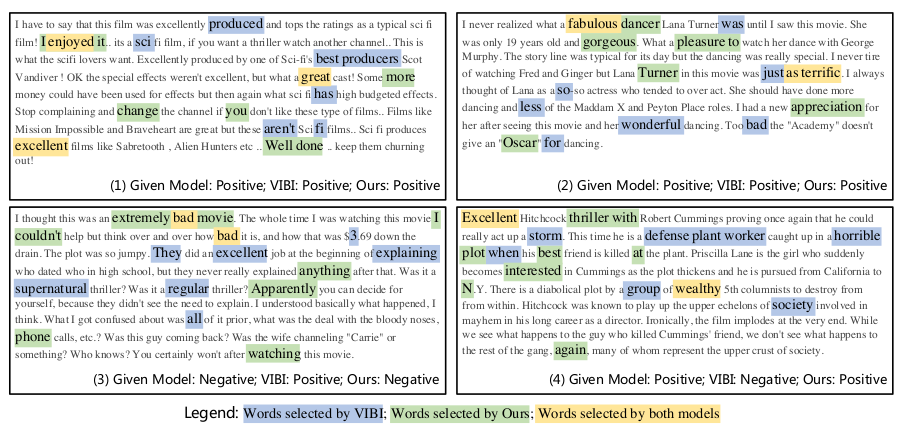
**评估结果**

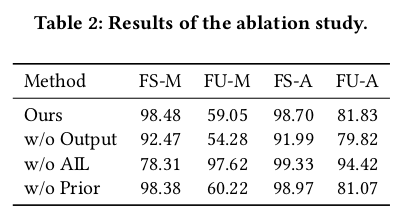


其他的模型没有GAN，所以FU-A和FU-M都较高。

MEED方法中，FU-M和FU-A会互相抑制。说明解释器E选择了一组合适的特征，未选中的特征是无用的且尽可能被拟合。

下图表明，MEED方法能在减少无倾向词的选择，能较为有效地减少歧义词的选择，以帮助解释方法做出正确的判断。

**消融研究**

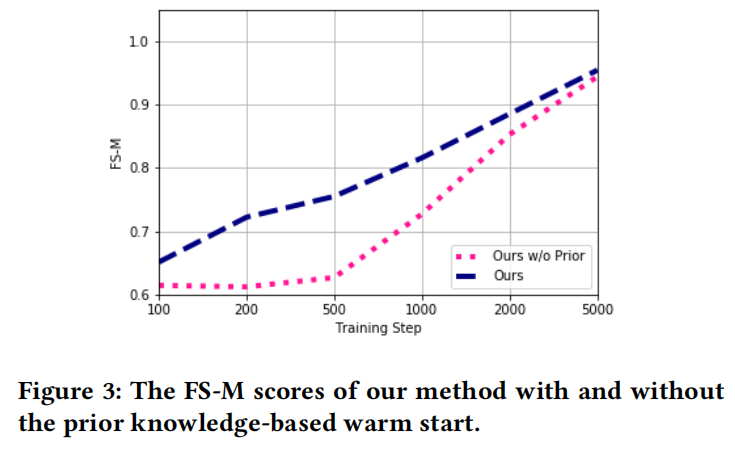


可以看到，去除AIL后FS-M下降，FU-M升高，这说明此时模型生成的解释的质量不高，证明了对抗学习（AIL）机制的有效性。同时也能观察到，使用原模型输出作为额外的监督信息的策略，和使用先验知识进行暖启动的策略并不能对保真度（Fidelity）指标有着显著影响。

**完备性（Sanity）检查**

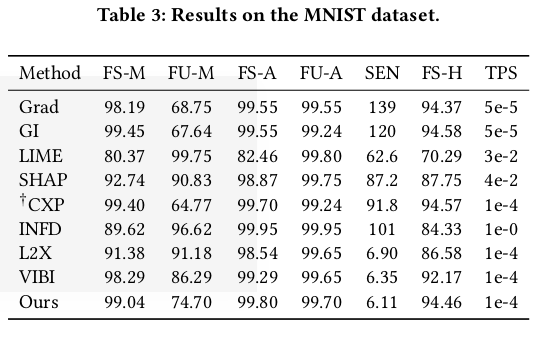
论文还使用了一种显著性检测的方法[3]对解释模型进行了完备性检查。其大致操作是将正常的生成的特征分数与数据随机化和黑盒模型参数随机化后生成的特征分数进行对比，二者得到的sanity score分别是9.39%和10.25%。这两个值越低代表着对数据和黑盒模型进行改动后生成的特征分数越不同，这表明这种解释方法是依赖于黑盒模型和数据本身的，解决了现有方法中的Sanity Problem。

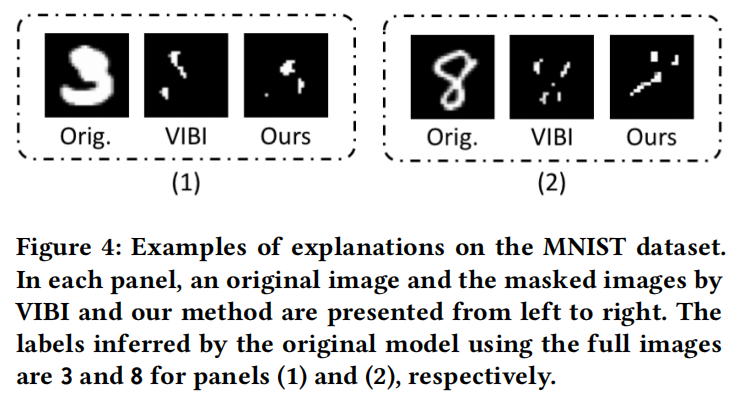
**暖启动（warm start）**



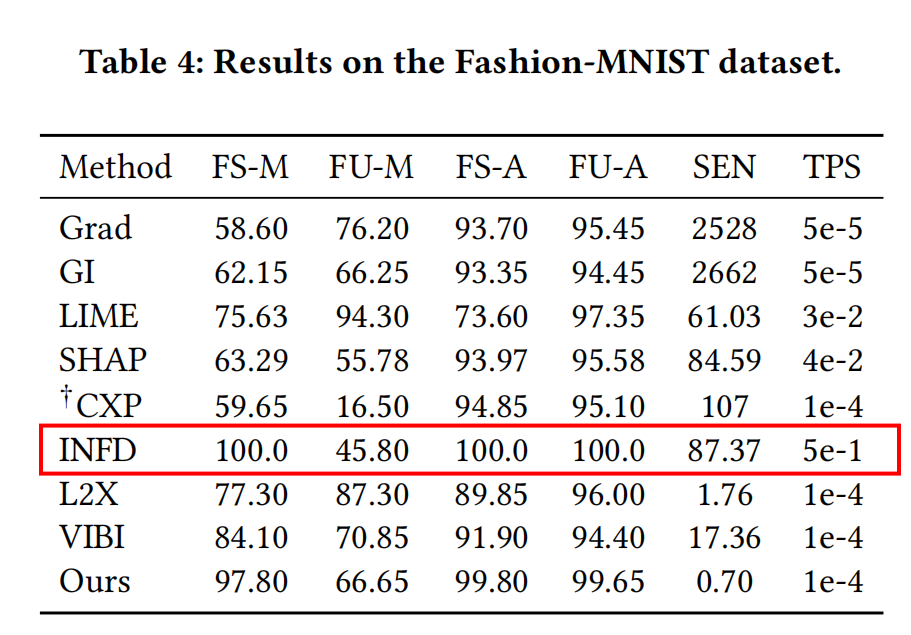
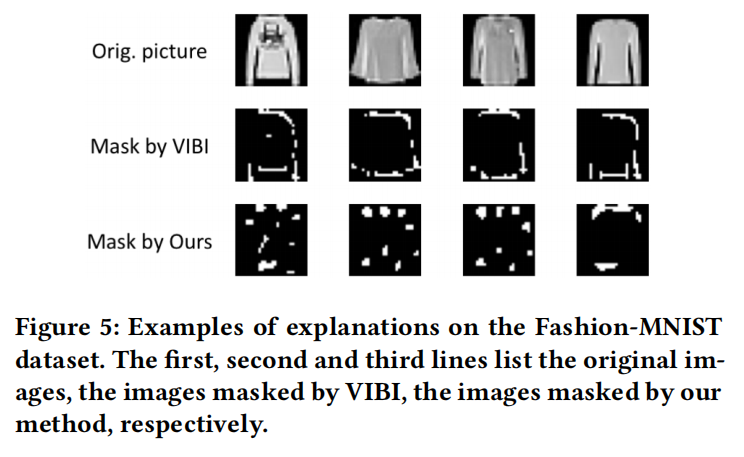
暖启动的效果如上图所示，虽然这一策略并不能对指标有显著影响，但可以有效地提高训练初期的收敛速度，这也代表着对抗学习中常出现的收敛困难的问题在一定程度得到改善。

**实验2/3：图像数据（不是最好的结果）**

**MNIST to classify 3 and 8**

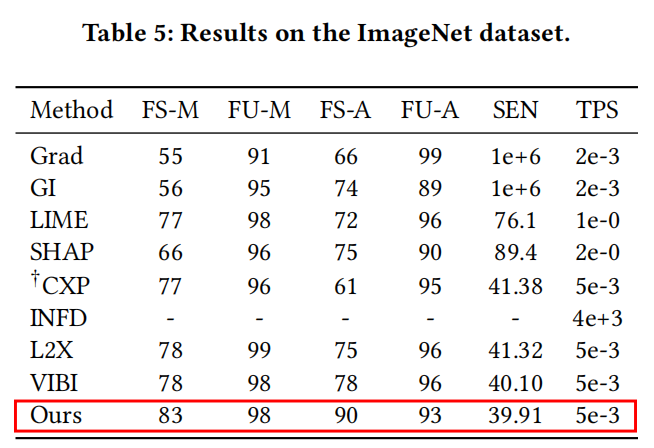
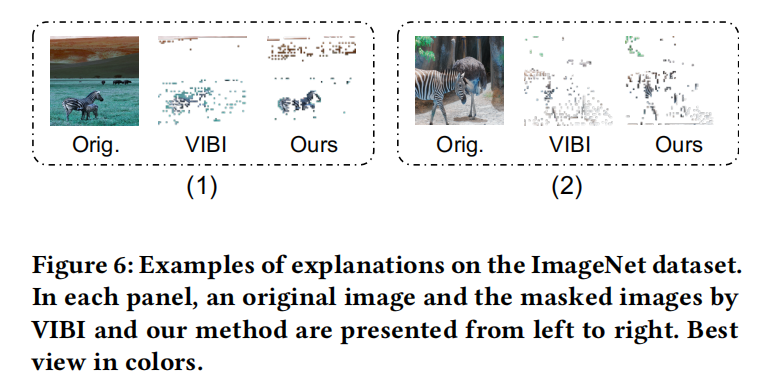
注意SEN的数据。在MNIST数据集上，MEED方法虽然不能达到最优的性能，但也能获得次优的效果。除此之外，MEED方法的SEN指标是最低的，这意味着该方法在保证性能的同时兼具良好的鲁棒性。  


**Fashion-MNIST to classify Pullover and Coat**

在Fashion-MNIST数据集上也能得到与MNIST数据集上相似的结论，因此不再赘述。

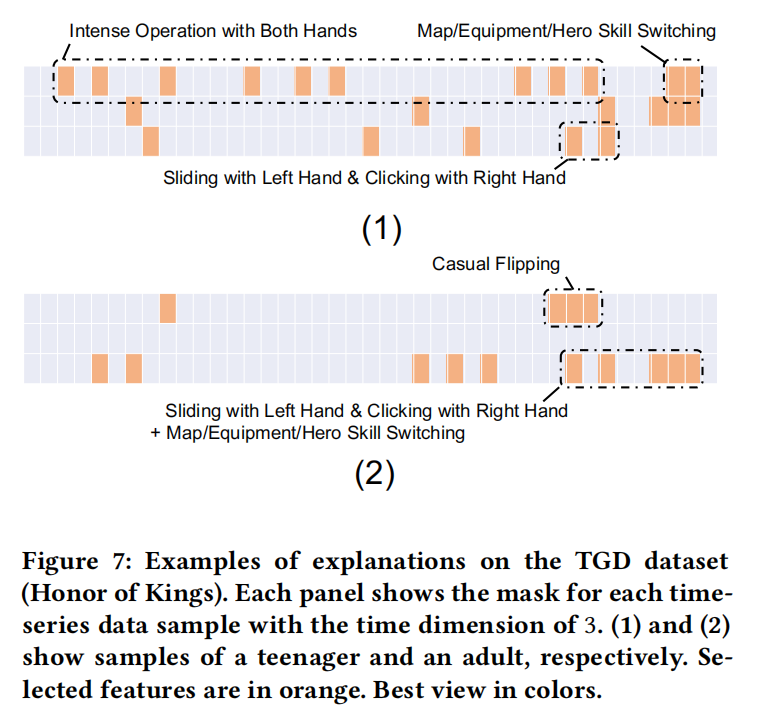
**ImageNet to classify Gorilla and Zebra**

在ImageNet上，MEED方法达到了最优的性能。除此之外，如给出的实例所示，该方法更多地关注于标签相关的区域，相比VIBI这一基线模型更具可解释性。

**实验3：时间序列数据**

**Tencent Honor of Kings gam for teenager recognition**



图中展示了王者荣耀中未成年人（1）和成年人（2）的操作序列数据，以及使用MEED方法选中的特征。MEED方法在该数据集上的指标 FS-M，FU-M，FS-A，FU-A， 和 SEN 分别是95.68%，82.24%，95.33%，82.37%，and 0.18%。从指标可以看出MEED方法构建出的预测模型有着很高的性能以及鲁棒性。除此之外，解释模型选中的特征也具有很强的可解释性。对于未成年人而言，选中的特征多集中于游戏前期；对于成年人而言，选中的特征多集中于游戏后期。这对应了游戏中两类人的行为模式，未成年人在游戏初期的操作比较复杂，越往后操作越单调；而成年人在游戏初期显得比较随意，但随着游戏进行，操作变得熟练且复杂。

## R Packages Used

arules. Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2021). arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-8. [https://CRAN.R-project.org/package=arules](https://cran.r-project.org/package=arules)

bookdown. Yihui Xie (2021). bookdown: Authoring Books and Technical Documents with R Markdown. R package version 0.24. [https://CRAN.R-project.org/package=bookdown](https://cran.r-project.org/package=bookdown)

Cairo. Simon Urbanek and Jeffrey Horner (2020). Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output. R package version 1.5-12.2. [https://CRAN.R-project.org/package=Cairo](https://cran.r-project.org/package=Cairo)

caret. Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-88. [https://CRAN.R-project.org/package=caret](https://cran.r-project.org/package=caret)

data.table. Matt Dowle and Arun Srinivasan (2021). data.table: Extension of data.frame. R package version 1.14.0. [https://CRAN.R-project.org/package=data.table](https://cran.r-project.org/package=data.table)

devtools. Hadley Wickham, Jim Hester and Winston Chang (2021). devtools: Tools to Make Developing R Packages Easier. R package version 2.4.2. [https://CRAN.R-project.org/package=devtools](https://cran.r-project.org/package=devtools)

dplyr. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. [https://CRAN.R-project.org/package=dplyr](https://cran.r-project.org/package=dplyr)

DT. Yihui Xie, Joe Cheng and Xianying Tan (2021). DT: A Wrapper of the JavaScript Library ‘DataTables’. R package version 0.19. [https://CRAN.R-project.org/package=DT](https://cran.r-project.org/package=DT)

e1071. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-8. [https://CRAN.R-project.org/package=e1071](https://cran.r-project.org/package=e1071)

ggplot2. Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani and Dewey Dunnington (2021). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.5. [https://CRAN.R-project.org/package=ggplot2](https://cran.r-project.org/package=ggplot2)

grid. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/).

gridExtra. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. [https://CRAN.R-project.org/package=gridExtra](https://cran.r-project.org/package=gridExtra)

gridExtra. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. [https://CRAN.R-project.org/package=gridExtra](https://cran.r-project.org/package=gridExtra)

iml. Christoph Molnar and Patrick Schratz (2020). iml: Interpretable Machine Learning. R package version 0.10.1. [https://CRAN.R-project.org/package=iml](https://cran.r-project.org/package=iml)

interactions. Jacob A. Long (2021). interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions. R package version 1.1.5. [https://CRAN.R-project.org/package=interactions](https://cran.r-project.org/package=interactions)

jpeg. Simon Urbanek (2021). jpeg: Read and write JPEG images. R package version 0.1-9. [https://CRAN.R-project.org/package=jpeg](https://cran.r-project.org/package=jpeg)

jtools. Jacob A. Long (2021). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.1.4. [https://CRAN.R-project.org/package=jtools](https://cran.r-project.org/package=jtools)

latex2exp. Stefano Meschiari (2021). latex2exp: Use LaTeX Expressions in Plots. R package version 0.5.0. [https://CRAN.R-project.org/package=latex2exp](https://cran.r-project.org/package=latex2exp)

lubridate. Vitalie Spinu, Garrett Grolemund and Hadley Wickham (2021). lubridate: Make Dealing with Dates a Little Easier. R package version 1.7.10. [https://CRAN.R-project.org/package=lubridate](https://cran.r-project.org/package=lubridate)

memoise. Hadley Wickham, Jim Hester, Winston Chang, Kirill Müller and Daniel Cook (2021). memoise: Memoisation of Functions. R package version 2.0.0. [https://CRAN.R-project.org/package=memoise](https://cran.r-project.org/package=memoise)

mgcv. Simon Wood (2021). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.8-36. [https://CRAN.R-project.org/package=mgcv](https://cran.r-project.org/package=mgcv)

mlbench. Friedrich Leisch and Evgenia Dimitriadou. (2021). mlbench: Machine Learning Benchmark Problems. R package version 2.1-3. [https://CRAN.R-project.org/package=mlbench](https://cran.r-project.org/package=mlbench)

mlr. Bernd Bischl, Michel Lang, Lars Kotthoff, Patrick Schratz, Julia Schiffner, Jakob Richter, Zachary Jones, Giuseppe Casalicchio and Mason Gallo (2021). mlr: Machine Learning in R. R package version 2.19.0. [https://CRAN.R-project.org/package=mlr](https://cran.r-project.org/package=mlr)

mmpf. Zachary Jones (2018). mmpf: Monte-Carlo Methods for Prediction Functions. R package version 0.0.5. [https://CRAN.R-project.org/package=mmpf](https://cran.r-project.org/package=mmpf)

numDeriv. Paul Gilbert and Ravi Varadhan (2019). numDeriv: Accurate Numerical Derivatives. R package version 2016.8-1.1. [https://CRAN.R-project.org/package=numDeriv](https://cran.r-project.org/package=numDeriv)

OneR. Holger von Jouanne-Diedrich (2017). OneR: One Rule Machine Learning Classification Algorithm with Enhancements. R package version 2.2. [https://CRAN.R-project.org/package=OneR](https://cran.r-project.org/package=OneR)

partykit. Torsten Hothorn and Achim Zeileis (2021). partykit: A Toolkit for Recursive Partytioning. R package version 1.2-15. [https://CRAN.R-project.org/package=partykit](https://cran.r-project.org/package=partykit)

patchwork. Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. [https://CRAN.R-project.org/package=patchwork](https://cran.r-project.org/package=patchwork)

png. Simon Urbanek (2013). png: Read and write PNG images. R package version 0.1-7. [https://CRAN.R-project.org/package=png](https://cran.r-project.org/package=png)

pre. Marjolein Fokkema and Benjamin Christoffersen (2021). pre: Prediction Rule Ensembles. R package version 1.0.1. [https://CRAN.R-project.org/package=pre](https://cran.r-project.org/package=pre)

R.utils. Henrik Bengtsson (2020). R.utils: Various Programming Utilities. R package version 2.10.1. [https://CRAN.R-project.org/package=R.utils](https://cran.r-project.org/package=R.utils)

randomForest. Fortran original by Leo Breiman, Adele Cutler, R port by Andy Liaw and Matthew Wiener. (2018). randomForest: Breiman and Cutler’s Random Forests for Classification and Regression. R package version 4.6-14. [https://CRAN.R-project.org/package=randomForest](https://cran.r-project.org/package=randomForest)

readr. Hadley Wickham and Jim Hester (2021). readr: Read Rectangular Text Data. R package version 2.0.1. [https://CRAN.R-project.org/package=readr](https://cran.r-project.org/package=readr)

rjson. Alex Couture-Beil (2018). rjson: JSON for R. R package version 0.2.20. [https://CRAN.R-project.org/package=rjson](https://cran.r-project.org/package=rjson)

roxygen2. Hadley Wickham, Peter Danenberg, Gábor Csárdi and Manuel Eugster (2021). roxygen2: In-Line Documentation for R. R package version 7.1.2. [https://CRAN.R-project.org/package=roxygen2](https://cran.r-project.org/package=roxygen2)

rpart. Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. [https://CRAN.R-project.org/package=rpart](https://cran.r-project.org/package=rpart)

RWeka. Kurt Hornik (2020). RWeka: R/Weka Interface. R package version 0.4-43. [https://CRAN.R-project.org/package=RWeka](https://cran.r-project.org/package=RWeka)

shiny. Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.6.0. [https://CRAN.R-project.org/package=shiny](https://cran.r-project.org/package=shiny)

svglite. Hadley Wickham, Lionel Henry, Thomas Lin Pedersen, T Jake Luciani, Matthieu Decorde and Vaudor Lise (2021). svglite: An ‘SVG’ Graphics Device. R package version 2.0.0. [https://CRAN.R-project.org/package=svglite](https://cran.r-project.org/package=svglite)

tidyr. Hadley Wickham (2021). tidyr: Tidy Messy Data. R package version 1.1.3. [https://CRAN.R-project.org/package=tidyr](https://cran.r-project.org/package=tidyr)

tm. Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. [https://CRAN.R-project.org/package=tm](https://cran.r-project.org/package=tm)

viridis. Simon Garnier (2021). viridis: Colorblind-Friendly Color Maps for R. R package version 0.6.1. [https://CRAN.R-project.org/package=viridis](https://cran.r-project.org/package=viridis)

xgboost. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng and Yutian Li (2021). xgboost: Extreme Gradient Boosting. R package version 1.4.1.1. [https://CRAN.R-project.org/package=xgboost](https://cran.r-project.org/package=xgboost)

yaImpute. Nicholas L. Crookston, Andrew O. Finley and John Coulston (2020). yaImpute: Nearest Neighbor Observation Imputation and Evaluation Tools. R package version 1.0-32. [https://CRAN.R-project.org/package=yaImpute](https://cran.r-project.org/package=yaImpute)

**可解释性网站**

https://christophm.github.io/interpretable-ml-book/r-packages-used.html

**论文目标与目的**

该工作是在模型可解释性领域内的研究，其提出了一个模型无关的IFS方法。其主要贡献在于提出了三种策略在一定程度上解决了现有IFS方法中存在的四个问题。该工作通过理论和大量实验证明了MEED方法在特征选择上的有效性和通用性，也证明了通过该方法选择的特征具有较高的质量。MEED方法在多种类型的数据集中均达到了SOTA的性能。

**关联关系**

1. AIL借鉴了GAN的对抗的思想，IFS的选择过程本身也可以看作是生成一个feature mask，这种对抗机制可以作用在局部以实现隐式的约束。
2. 在引文注意到一篇发表在NAACL 2019的《Attention is not Explanation》[2]。Attention和IFS表面上看都是一种分配权重的机制，只是二者的目的不同。虽然Attention并不一定具备可解释性，但也许可以结合IFS和Attention的共通之处去指导去建立一个本身就具有可解释性的复杂网络。而不是需要依赖一些黑盒解释方法。用黑盒解释黑盒是需要比较严谨的推导的，没有经过严谨推导的解释模型只能给予有限的信任，一个例子是发表在NIPS 2018的《Sanity Checks for Saliency Maps》[3]就证明了一些广泛使用的saliency method是独立于训练数据和模型，这会导致在某些任务上的失效。

**相关资料：**

[1] <https://www.jiqizhixin.com/articles/2019-10-30-9>  
[2] Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3543–3556.  
[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. 9505–9515.