多模态模型可解释性

王鹏飞

论文列表

1. ICCV2021\_Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers

2. AAAI2022\_Evaluating Explainable AI on a Multi-Modal Medical Imaging Task Can Existing Algorithms Fulfill Clinical Requirements

|  |
| --- |
| ICCV2021  [Hila Chefer](https://hila-chefer.github.io/) [Shir Gur](https://www.gurshir.com/) [Lior Wolf](http://www.cs.tau.ac.il/~wolf/)  [homepage](https://hila-chefer.github.io/publication/generic-attention-model-explainability/) |
| 1. **研究背景** 2. 任务   对基于Bi-Modal 和Encoder-Decoder 的transformer模型进行解释。   1. 挑战  * 使用 co-attention 的transformer结构需要**并行考虑多个attention maps**，进而**突出显示**与*模型预测相关*的**输入侧信息**。 * Co-attention和self-attention的异同Figure1，queries来源于target domain和keys和values来源于 context domain  1. 贡献  * 本文第一个提出可解释多种形式的attention，包括self-attention，co-attention以及encoder-decoder attention，可用于跨模态的解释工作。 |
| 1. **整体框架** 2. 问题定义：  * M是 基于transformer的模型 * 模型输入是序列，其中的一个或者两个，解释性即高亮出出，中决定模型M输出的tokens。   例1：是一句文本描述，是一张图像, 模型（CLIP）    例2: 是一张图像，模型DETR（目标检测）     1. Transformer- based 模型 |
| 1. **方法** |
| 本文主要定义四个relevance矩阵，用来计算input tokens 之间的关系，并且通过更新relevancy map来表示transformer中attention模块的显著性。   1. Relevance初始化      1. Relevance更新  * 对于Self-attention     利用用attention map A和梯度来更新relevancy maps R       * 对于co-attention   两个流或者分支进行归一化，平衡不同特征的相关性。    Bi-model 再进行更新     1. **实验** |
| 1. 正-负扰动测试   根据产生的relevance map获得token的重要性，逐步删除token并测试top-1 accuracy。  对于正扰动测试，重要性从高到低删除token；负扰动则相反，从低到高删除。如果求得的relevance map比较准确，正扰动删除token后accuracy极速下降；相反，负扰动删除的token对分类accuracy影响不大。  两种测试均使用AUC来评估，另外对于两种模态分别做正-负扰动测试。   * VisualBERT是仅使用self-attention网络结构     本文方法优于sota，并且在text 扰动上远超归因transformer的方法。   * LXMERT，使用self-attention and co-attention网络结构。     对于image token负扰动，可发现删除接近80% ，accuracy还未明显降低；对于正扰动，发现accuracy出现骤降。说明本文方法学习到的relevancem重要性准确。   1. token重要性可视化     和基准方法比较，文本结果比较相似，但图像的显著区域本文方法关注地更集中。   1. 产生的重要区域和gt比较。      * DETR，encoder-decoder model     产生的显著区域和检测框重合度比较高。 |
| 1. **结论**  * 对于跨模态任务Co-Attention进行了比较好的解释。 * 证明模态之间有信息互补的特性。 * 本文方法比较好的跟踪注意力图的演变和混合。 |

|  |
| --- |
| AAAI2022  [Weina Jin](https://weina.me/)  [Xiaoxiao Li](https://xxlya.github.io/xiaoxiao/) [Ghassan Hamarneh](https://www.sfu.ca/computing/people/faculty/ghassanhamarneh.html) |
| **一、研究背景**  任务：   1. 提出两个评测指标， 评估多模态医学图像可解释性方法。   挑战：   1. 由于每种模态对潜在相同的生物医学现象具有不同的临床意义，热力图方法可能无法满足。 2. 模态相关的特征对于临床用户理解AI决策至关重要，现行方法忽略。   贡献：   1. 对医学成像任务进行系统评估，包括定量和定性医学评测，满足临床要求的忠实性和合理性解释。 2. 对16种现行三大类activation-，gradient-， perturbation评估。 3. 提出的指标（MSFI）能够综合考虑模态优先级和特征定位的临床模式来自动化人类评估过程。 |
| **二、整体框架**    主要计算MI和MSFI两个指标，MI衡量每个模态的重要性，MSFI评测每个模态中病兆区域的定位。 |
| **三、方法**   1. Modality Importance (MI)   计算出估计出来的每个模态重要性和目标shapely值之间的差异性。   * 把m个模态当成合作博弈论的一个参与者，计算shapely值做为gt-MI。     其中（1）中v是一个模态相关的表现指标，在测试集上计算模型的accuracy；将子集中未包含的模态中的所有值设置为0；零消融实验中由特征定位图定义的局部特征区域去除，而不是整体的一个模态都去除。   * MI correlation   计算当前解释性方法的MI和 上一步gt-MI的相关性。  估计MI的计算方法：把每一个模态产生的saliency map中所有正数相加求和。  相关性计算方法：Kendall’s Tau-b ，求出来的范围[-1,1]      参考维基百科: [Kendall rank correlation coefficient](https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient#Tau-b)   1. Modality-Specific Feature Importance (MSFI)   计算出病兆区域和标注的mask之间的差异性。 |
| 1. **实验** 2. 实验准备  * 任务   对胶质瘤进行分类，high-grade gliomas和lower-grade gliomas   * 数据集   BraTS 2020 和 用gan 合成的BraTS-based synthetic， 其中合成的数据病兆区域的形状和大小更可控，每个数据集都有四个模态。   * 模型学习   VGG-like 3D, DenseNet121，多模态在模型输入端融合。   * 评测的方法，主要是三大类     2、 评测结果   * MI     大多数算法通常可以正确反映模型决策的重要模式，但个别数据点的变化较大，除（GradCAM、KernalSHAP、Feature Permutation）只能为所有模式生成一个相同的热力图，即为NaN。   * MSFI       为了测试MSFI测量的heatmap是否可以作为模型决策质量的指标，我们将数据分为正确或错误的预测组。使用Mann-Whitney U test衡量是否有显著差异。   * Non-Modality-Specific     文中还对非模态相关的一些指标做了对比，蓝色是忠实度，黄色合理性。  忠实度：   * DiffAUC通过计算算法与其基线（随机消融）之间的曲线下面积差（AUC）来量化性能恶化的程度   合理性：   * IOU   IOU =  计算显著性找出的病兆区域A和GT-MASK B的交集和并集商   * Feature Portion (FP, the sum of heatmap values inside the ground-truth feature mask over the total values).)   FP =   * Doctor’s rating     可以发现：   * 一些基于梯度的算法优于基于激活或扰动的算法。 * MSFI 是FP的泛化形式 |
| 1. **讨论**  * 现行的可解释性方法无法满足临床，最好方法的MSFI 平均值也仅在0.5左右。   + 医生倾向于认为解释性完全忠实于模型的决策过程，因此会通过判断解释的合理性来接受或拒绝模型的建议。   + 热力度变化的原因也需要进一步研究。 * MSFI可代替医生手动评估XAI算法是否能够满足多模态医学图像设置的临床要求 |
| 1. **总结**  * 优势 * 提出的MSFI能够编码出*模态优先级*和*特征定位，*这两个指标满足临床需求。 * 对16种可产生热力图的解释性方法进行了比较好的评价，并指出都不能满足临床需要。 * 未来把这指标引入到解释性方法的学习中，能够反映出模型的真实表现和临床的先验知识。 * 不足 * 对非产生热力图的可解释性方法没有进行评价和比较。 |