Indian Institute of Information Technology Kottayam

Advancing Trust and Explainability in Artificial Intelligence Systems

by

Harsh Vardhan Singh Chauhan

2022BCD0044

Under guidance of

Dr Jeena Thomas

Contents

Introduction

- **AI Models as Black Boxes** – Large models like transformers lack transparency, making decisions hard to interpret.

- **Trust & Transparency** – Explainability ensures users and stakeholders can trust AI predictions. (healthcare, finance etc.)

- **Challenges** – Trade-off between accuracy and interpretability; scaling explainability to large models is difficult.

Literature Survey

Ref No	Method	Category	Dataset	Model	Task	Approach
[1]	Grad CAM	Common Attribution	ImageNet, PASCAL VOC, COCO, VQA	VGG-16, ResNet, GoogleNet, AlexNet	Classification, Captioning, VQA	Gradient based activation mapping
[2]	Integrated Gradients	Common Attribution	ImageNet	GoogleNet	Classification	Axiomatic attribution
[3]	ViT Shapley	Common Attribution	ImageNette, MURA, Oxford IIIT-Pets	Vision Transformers	Classification	Shapley value estimation for feature importance
[4]	LRP	Common Attribution	CIFAR-10, ImageNet, MIT Places	CNNs	Classification	Taylor expansion for feature attribution

Literature Survey

Ref No	Method	Category	Dataset	Model	Task	Approach
[5]	Raw Attention	Attention based	Crop Satellite Data	LSTM-RNN, MS-ResNet, TempCNN	Classification,	Raw self-attention for time-series
[6]	Attention Rollout & Flow	Attention based	Subject-verb agreement Dataset	BERT	Sentiment analysis	Graph-based quantification of attention flow in transformers
[7]	Grad-SAM	Attention based	Stanford Sentiment Tree, AgNews, IMDB, MultiRC	BERT based models	Sentiment Analysis	Uses gradient with self-attention for activation maps
[8]	Beyond Attention	Attention based	ImageNet, Movie Review, ERASER	BERT, Vision Transformer	NLP & Vision Tasks	Combines attention and propagation

Literature Survey

Ref No	Method	Category	Dataset	Model	Task	Approach
[9]	Vision DiffMask	Pruning based	CIFAR-10, ImageNet	LSTM-RNN, MS-ResNet, TempCNN	Classification	Differentiable patch masking for hidden layer activations
[10]	X-Pruner	Pruning based	CIFAR-10, ILSVRC-12	BERT, Swin Transformer	Classification	Differentiable masks for unit contribution, layer-wise pruning
[11]	EViT	Pruning based	ImageNet, JFT-300M	Transformers	Classification	Identifies and fuses inattentive tokens
[12]	IA-Red2	Pruning based	ImageNet, Kinetics-400	Vision Transformers	Classification	Policy based dropout

Literature Survey

Ref No	Method	Category	Dataset	Model	Task	Approach
[13]	ViT-CX	Inherently Explainable	ImageNette, MURA	Vision Transformers	Classification	Causal explanation using feature maps and clustering masks
[14]	ViT-NeT	Inherently Explainable	CUB-200-2011, Stanford Cars, Stanford Dogs	Vision Transformers	Classification	Neural tree-based decoder
[15]	R-Cut	Inherently Explainable	ImageNet, LRN	Vision Transformers	Classification	Relationship-weighted explanation and token cutting
[16]	eX-ViT	other	PASCAL VOC 2012, MS COCO 2014	Vision Transformers	Weakly Supervised Segmentation	Explainable Multi-Head Attention & Attribute-guided Explainer

Problem Statement

- **Limited Vision-based tasks**– Most explainability methods focus only on **vision classification**, ignoring multi-modal understanding

- **Single-Model Focus** – Existing methods work with either **CNNs or Transformers**, but rarely both

- **Lack of Multi-Modal Explainability** – Current approaches fail to **link textual components with specific image regions** in vision-language models.

Objectives

- **Develop a Novel Explainability Framework** – Design an approach that works for both **CNNs and Transformer-based models**.

- **Enable Multi-Modal Interpretation** – Provide insights into **text-image correlations** for tasks like **VQA and image captioning**.

- **Go Beyond Classification** – Extend explainability to **retrieval, segmentation, and multi-modal reasoning** tasks.

Future Work

- Exploration of **multi-modality explainability** framework.

- **Strength and drawbacks** for explainable approach and methodologies.

- Designing **novel architecture** for the multi modal systems (transformer, CNN etc.)

References

- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). **[1]**
- Sundararajan, M., Taly, A. and Yan, Q., 2017, July. Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR. **[2]**
- Covert, I., Kim, C. and Lee, S.I., 2022. Learning to estimate shapley values with vision transformers. *arXiv preprint arXiv:2206.05282*. **[3]**
- Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R. and Samek, W., 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25* (pp. 63-71). Springer International Publishing. **[4]**

References

- Rußwurm, M. and Körner, M., 2020. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, *169*, pp.421-435. **[5]**
- Abnar, S. and Zuidema, W., 2022. Quantifying attention flow in transformers. arXiv 2020. *arXiv preprint arXiv:2005.00928*. **[6]**
- Hauon, E., 2023. *Grad-SAM: Explaining Transformers via Gradient Self-Attention Maps* (Master's thesis, Reichman University (Israel)). **[7]**
- Chefer, H., Gur, S. and Wolf, L., 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 782-791). **[8]**

References

- Nalmpantis, A., Panagiotopoulos, A., Gkountouras, J., Papakostas, K. and Aziz, W., 2023. Vision diffmask: Faithful interpretation of vision transformers with differentiable patch masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3756-3763). **[9]**
- Yu, L. and Xiang, W., 2023. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24355-24363). **[10]**
- Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J. and Xie, P., 2022. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*. **[11]**

References

- Pan, B., Panda, R., Jiang, Y., Wang, Z., Feris, R. and Oliva, A., 2021. IA-RED $^ 2$: Interpretability-Aware Redundancy Reduction for Vision Transformers. *Advances in Neural Information Processing Systems*, *34*, pp.24898-24911. **[12]**
- Xie, W., Li, X.H., Cao, C.C. and Zhang, N.L., 2022. Vit-cx: Causal explanation of vision transformers. *arXiv preprint arXiv:2211.03064.* **[13]**
- Kim, S., Nam, J. and Ko, B.C., 2022, June. Vit-net: Interpretable vision transformers with neural tree decoder. In *International conference on machine learning* (pp. 11162-11172). PMLR. **[14]**

References

- Niu, Y., Ding, M., Ge, M., Karlsson, R., Zhang, Y., Carballo, A. and Takeda, K., 2024. R-cut: Enhancing explainability in vision transformers with relationship weighted out and cut. *Sensors*, *24*(9), p.2695. **[15]**
- Yu, L., Xiang, W., Fang, J., Chen, Y.P.P. and Chi, L., 2023. ex-vit: A novel explainable vision transformer for weakly supervised semantic segmentation. *Pattern Recognition*, *142*, p.109666. **[16]**

Thank you